2009-11-25

Electromagnetic Scattering Solutions for Digital Signal Processing

Jonathan Blackledge
Technological University Dublin, jonathan.blackledge@tudublin.ie

Recommended Citation

Blackledge, J.:Electromagnetic Scattering Solutions for Digital Signal Processing. Jyvaskyla Studies in Computing 108;Department of Mathematical Information Technology, University of Jyvaskyla, 2009, pp.1-296 ISBN 978-951-39-3741-6, issue: ISSN 1456-5390

Jonathan Blackledge

Electromagnetic Scattering Solutions for Digital Signal Processing

November 25, 2009

PhD Thesis: Studies in Computing Series 108

UNIVERSITY OF JYVÄSKYLÄ 2009

ABSTRACT

Blackledge, Jonathan
Electromagnetic Scattering Solutions for Digital Signal Processing
Jyväskylä: University of Jyväskylä, 2009, 297 p.
(Jyväskylä Studies in Computing
ISSN 1456-5390; 108)
ISBN 978-951-39-3741-6

Electromagnetic scattering theory is fundamental to understanding the interaction between electromagnetic waves and inhomogeneous dielectric materials. The theory unpins the engineering of electromagnetic imaging systems over a broad range of frequencies, from optics to radio and microwave imaging, for example. Developing accurate scattering models is particularly important in the field of image understanding and the interpretation of electromagnetic signals generated by scattering events. To this end there are a number of approaches that can be taken. For relatively simple geometric configurations, approximation methods are used to develop a transformation from the object plane (where scattering events take place) to the image plane (where a record of some measure of the scattered field is taken). The most common approximation is the weak scattering approximation which ignores the effect of multiple scattering interactions and the first part of this thesis investigates the use of this approximation for electromagnetic imaging systems modelling. When scattering interactions become progressively more complex (e.g. multiple scattering from random media), the applications of deterministic scattering theory becomes difficult to use in practice. Consequently the inverse scattering problem can become ill-posed. For this reason, a number of other approaches are considered which include developing statistical models for the scattered field itself rather than the scatterer. In this thesis, we investigate the use of diffusion based models for solving the inverse scattering problem when strong scattering processes occur. We then extend this approach and consider the intermediate case by modelling the scattering processes using a fractional diffusion equation. Finally, a low frequency scattering theory is presented which leads to the proposition that light and other high frequency electromagnetic wavefields can be weakly diffracted by a low frequency scattered field. This leads to a new interpretation of gravity gravitational lensing which is investigated through the question as to why Einstein rings, observed in the visible spectrum, are blue?

Keywords: Electromagnetic fields and waves, scattering theory, inverse scattering solutions, exact inverse scattering theory, scattering from random media, diffusion based models, fractional diffusion, fractaional calculus, intermediate scattering models, low frequency scattering.

Author

Jonathan Blackledge
jonathan.blackledge@dit.ie
Department of Mathematical Information Technology,
Faculty of Information Technology,
University of Jyväskylä,
Finland

Supervisors

Professor Timo Hämäläinen
timo.t.hamalainen@jyu.fi
and
Professor Jyrki Joutsensalo
jyrki.j.joutsensalo@jyu.fi
Department of Mathematical Information Technology,
Faculty of Information Technology,
University of Jyväskylä,
Finland

Reviewers

Prof Andrey Garnaev
garnaev@yahoo.com
Department of Computer Modelling and Microprocessor Systems,
Faculty of Applied Mathematics and Control Processes,
St Petersburg State University, Russia.

Dr Peter Sherar
p.sherar@cranfield.ac.uk
Applied Mathematics and Computing Group,
Faculty of Engineering,
Cranfield University, England.

Dr Ardie Osanlou
Ardie.o@optictechnium.com
Centre for Modern Optics,
OpTIC Technium OpTIC,
St Asaph Business Park,Wales

Opponent

Prof Michael Rycroft
michael.rycroft@btinternet.com
Director of Research,
International Space University,
Strasbourg, France

ACKNOWLEDGEMENTS

The author acknowledges the support of the School of Electrical Engineering Systems, Faculty of Engineering, Dublin Institute of Technology and the Science Foundation Ireland who are supporting the author through the Stokes Professorship Programme. The author also acknowledges the help and advice of Professor Eugene Coyle (Head of the School of Electrical Engineering Systems) and Dr Marek Rebow (Head of Research for the Faculty of Engineering) at Dublin Institute of Technology and Professors Timo Hämäläinen and Jyrki Joutsensalo, Department of Mathematical Information Technology, University of Jyväskylä, Finland.

GLOSSARY

Alphabetical

$A(\mathbf{k})$	Amplitude spectrum
$\mathbf{A}$	Magnetic vector potential
$\mathbf{b}$	Microscopic magnetic field
$\mathbf{B}$	Magnetic field density
c	Wavespeed
c_0	Wavespeed of free space (e.g. speed of light)
D	Fractal dimension, scale size of an object or Diffusivity
$\mathbf{D}$	Electric displacement
1D	One-Dimensional
2D	Two-Dimensional
3D	Three-Dimensional
$\mathbf{e}$	Microscopic electric field
$\mathbf{E}$	Macroscopic electric field
$f(\mathbf{r})$	Arbitrary real (or complex) function (typically the object function or system input)
$\mid f(\mathbf{r}) \mid$	modulus of complex variable or function f
$\|f(\mathbf{r})\|$	Norm (e.g. a Euclidean L_2-norm) of a continuous function f
f_{ij}	2D discrete function (in real space)
$\|f_{ij}\|$	Norm (e.g. a Euclidean ℓ_2-norm) of a discrete function (e.g. 2D array or matrix) f_{ij}
$F(\mathbf{k})$	Complex spectrum of function $f(\mathbf{r})$
F_r	Real component of spectrum
F_i	Imaginary component of spectrum
$g(\mathbf{r} \mid \mathbf{r}_0, k)$	Time independent Green function
$G(\mathbf{r} \mid \mathbf{r}_0, t \mid t_0)$	Time dependent Green function
$\mathrm{Im}[f]$	Imaginary part of complex variable or function f
H	Hurst exponent
$\mathbf{H}$	Macroscopic magnetic field
$\mathcal{H}$	Hausdorff space
$\mathbf{j}$	Charge density
k	Wavenumber ($= 2\pi/\lambda$)
k_x	Spacial frequency in the x-direction
k_y	Spacial frequency in the y-direction
$\mathbf{k}$	Wave vector $= \hat{\mathbf{x}}k_x + \hat{\mathbf{y}}k_y$
$n(\mathbf{r})$	Noise function
$\hat{\mathbf{n}}$	Unit vector
$N(\mathbf{k})$	Noise spectrum
$O(x, y)$	Object function
$\widetilde{O}$	Fourier transform of object function

$p(\mathbf{r})$	Instrument function, or Point Spread Function
$P(\mathbf{k})$	Optical Transfer Function [Fourier transform of $p(\mathbf{r})$]
P_{ij}	Discrete Optical Transfer Function (DFT of p_{ij})
$\mathrm{Pr}[x(t)]$	Probability density Function
$P(\mathbf{k})$	Power spectrum ($=\mid F(\mathbf{k}) \mid^2$)
q	Fourier dimension
$\mathrm{Re}[f]$	Real part of complex variable or function f
$\mathbf{r}$	General position vector in a 2D or 3D space (depending on the context)
$d^2\mathbf{r}$	Surface element $dxdy$
$d^3\mathbf{r}$	Volume element $dxdydz$
$s(\mathbf{r})$	Real or complex (analytic) image
S	Surface
$\mathrm{sinc}(x)$	Sinc function ($= \sin(x)/x$)
t	Time
$u(\mathbf{r},t)$	General scalar wavefield function)
u_i	Incident wavefield
u_s	Scattered wavefield
V	Volume
x,y,z	General independent variables
z_0	Free space wave impedance
$\in$	In (e.g. $x \in [a,b)$ is equivalent to $a \leq x < b$)
$\forall$	Forall (e.g. $f(t) = 0, \forall t \in (a,b]$)

Greek

α	Chirping parameter
γ	General scattering function
$\Gamma(q)$	Gamma function $= \int_0^\infty x^{q-1}e^{-x}dx$
δ^n	n-dimensional Dirac delta function
ϵ	Permittivity
ϵ_0	Permittivity of free space
θ	Phase, angle
λ	Wavelength
σ	Conductivity
μ	Permeability
μ_0	Permeability of free space
ρ	Charge density, or material density
ω	Angular frequency

Operators

$\hat{D}$	Homogeneous linear differential operator
$\mathcal{F}_1$	One dimensional Fourier transform
$\mathcal{F}_1^{-1}$	One dimensional inverse Fourier transform
$\mathcal{F}_2$	Two dimensional Fourier transform
$\mathcal{F}_2^{-1}$	Two dimensional inverse Fourier transform
$\mathcal{H}$	Hilbert transform
$\hat{L}$	Inhomogeneous linear differential operator
$\otimes_n$	n-dimensional convolution operation causal or otherwise (depending on the context specified)
$\otimes\otimes$	$\equiv \otimes_2$
$\otimes$	$\equiv \otimes_1$ or $\equiv \otimes_2$ or $\equiv \otimes_3$ depending on context, i.e. dimension of functions
$\odot_n$	n-dimensional correlation operation - continuous or discrete, causal or otherwise (depending on the context specified)
$\odot\odot$	$\equiv \odot_2$
$\odot$	$\equiv \odot_1$ or $\equiv \odot_2$ or $\equiv \odot_3$ depending on context, i.e. dimension of functions
$\Longleftrightarrow$	Transformation into Fourier space
$\longleftrightarrow$	Transformation into some transform space (as defined)
∇^2	Laplacian operator

LIST OF FIGURES

LIST OF TABLES

CONTENTS

1 INTRODUCTION

Scattering theory is very important due to the wide range of applications in which it must be applied to interpret the characteristics of signals and images. Applications are varied and include radar sensing, medical ultrasound, semiconductor wafer inspection, polymerization process monitoring, acoustic tiling, free-space communications, computer-generated imagery and so on. Major research problems in scattering theory often involve predicting how various systems will scatter radiation. A widely studied but more difficult challenge is the inverse scattering problem, in which the goal is to observe scattered radiation and use that observation to determine properties of either the scatterer or the radiation field pattern before scattering has occurred. In general, the inverse problem is not unique. Different types of scatterers can give rise to the same pattern of scattered radiation and thus, the problem is not solvable in the general case. For this reason, developing new approaches and methods for improving the extraction of information from signals and image obtained by recording a scattered field is a very important and topical field of research.

Electromagnetic (EM) scattering theory is fundamental to modelling the interaction of EM waves with matter. This has been an important topic for many years as the 'physics' associated with EM signals and images and the engineering that is applied to develop different EM 'imaging systems' is usually based on some type of scattering model. In turn, information retrieval from data produced by such systems relies on the design of data processing and data analysis software that is often, in effect, based on solutions to the inverse scattering problem. The details and complexity of the algorithms designed for this purpose vary from one application to the next but at the centre of any application is a model based on the scattering of an EM wavefield from an inhomogeneous (dielectric) material. The goal of this thesis has been to develop the mathematical techniques used for modelling EM imaging systems and show how such models can be used as a guide to the interpretation of the data captured by different imaging systems from which suitable image processing algorithms can be designed. An underlying theme is the relationship between the development of image processing systems and the 'physics' of waves and vibrations. Central to this relationship is the role of scattering theory and, in particular, the Green's function solution to an appropriate linear inhomogeneous wave equation

that can be taken to describe (usually to a limited extent) the 'physics' associated with a particular type of wavefield.

1.1 Scattering Theory and the Imaging Equation

Conventional models used for developing a basic 'imaging equation' (i.e. an equation that models the relationship between the object and image planes) and their relationship to the properties of a detected scattered wavefield are almost exclusively based on application of approximation methods used to model a scattered field that is measured in the far-field, i.e. along way from the position in space where the scattering interactions have occurred. The most common of these approximation methods is the weak scattering approximation. The weak scattering approximation assumes that the scattered field is the result of single scattering processes only. When this approximation is considered in the far-field, a simple mapping is obtained between the scattered field and the scattering function that is compounded in the Fourier transform. This result is the basis for Fourier optics, for example, the essential point being that, in the focal (or Fourier plane), a lens can be taken to perform a Fourier transform of the input, i.e. the light wavefront that is normally incident on the back of the lens. In this case, the scattering function is glass and a wavefront is taken to be a plane wavefield. In the intermediate field, when a wavefront is taken to have a parabolic curvature, the Fourier transform is replaced with the Fresnel transform which is characterised by a quadratic phase factor. Although the geometry associated with the Fresnel transform is different to that of the Fourier transform, both transforms are based on the same weak scattering approximation.

The weak scattering approximation is often referred to as the Born approximation after Max Born, who first considered the approximation with regard to scattering processes in quantum mechanics through solutions to the Schrödinger equation. This requires that the 'scattering model' adheres to the 'weak field' condition in which the total scattered field is considered to be a weak perturbation of the incident field in terms of some appropriate measure. In turn, depending on the complexity of the scattering model, this condition can usually be quantified in terms of physical parameters such as the wavelength λ of the incident wavefield and the scale length L of the scatterer, a basic 'standard' being that $\lambda >> L$. The problem with this condition is that it is fundamentally incompatible with a basic requirement associated with systems that are designed to recover information at a resolution compatible with the scale of the wavelength, i.e. when $\lambda \sim L$. Thus, any system that is designed and engineered to 'image' an object in some way on the scale of the wavelength of the incident field is prone to distortion due to the effects of multiple scattering, an effect that is not incorporated within the weak scattering condition. Instead, multiple scattering processes are considered to contribute to the noise function of the system. One of the principal aims of this thesis has been to investigate multiple scattering models and solutions that can be used to process EM signals and images. To do this, it is necessary to combine scattering theory with the

engineering principles and fundamental models associated with imaging systems.

1.2 Imaging Systems

All imaging systems can be viewed in terms of some appropriate instrument that, by default, is only able to record a scattered field to a limited extent. The relationship between the object plane and the image plane is determined by the 'instrument function' that, in turn, either directly or indirectly (i.e. after appropriate data processing has been applied), determines the characteristics of the image via the Point Spread Function. The fundamental imaging equation is given by

$$s(x,y) = p(x,y) \otimes_2 f(x,y) + n(x,y) \tag{1.1}$$

where $\otimes_2$ denotes the two-dimensional convolution integral, f is the object function (a description of the object plane), p is the Point Spread Function, n is the noise (function) and s is the signal (one-dimensional case) or image (as in the two-dimensional given here) or a three dimensional image, depending upon the application.

The Point Spread Function (PSF) is taken to be invariant of different positions in the image plane and the process is therefore stationary or 'isoplanatic'[1]. This allows the convolution theorem to be applied providing a route to the analysis and processing of an image in Fourier space using the Fourier transform. However, if the PSF varies in the image plane, the convolution process is non-stationary and the convolution theorem cannot be applied in the same way. This has important consequences for developing methods involved in solving the fundamental inverse problem: given s, p and a statistical model for n (i.e. the Probability Density Function for n), find f. For the stationary case, Fourier based methods can be used to design a range of (inverse) filters (e.g. deconvolution algorithms) but, for non-stationary problems, the filters must designed be applied algebraically. This may involve solving large systems of linear equations of size $n^2 \times m^2$ for digital images of size $n \times m$ leading to problems of numerical stability. A historically important case of this problem occurred when the Hubble Space Telescope was first launched and it was found that minor errors in the curvature of the primary reflector were leading to blurred images that were non-isoplanatic. Numerical based compensation of these effects were considered impractical and compensating optics were designed and implemented instead.

Under ideal circumstances, by accurately modelling an imaging system, it is possible to derive a description for the relationship between the object and image planes, identify the nature of the inverse problem and thus, develop an appropriate reconstruction method as required. However, the accuracy of the model has to be balanced with the simplicity of the results that can be derived from it in terms of designing algorithms that are of practical and 'engineering' value. Achieving the right balance is central to imaging systems modelling and image understanding. A

[1] A term that is used primarily in the field of Optics

key feature to achieving this balance, which is fundamental to all scatter imaging systems, is that the physical effects of strong scattering together with other incompatibilities and errors associated with the representation of a recorded image in terms of equation (1.1) are combined to form the noise term. In other words, equation (1.1) is based on the weak scattering approximation where $\lambda >> L$ and is therefore incompatible with the criterion that the information content of an image is based on interactions that take place on the scale of the wavelength (i.e. λ L). Thus, although equation (1.1) is used as a fundamental model for an image, it is generally incompatible with the weak scattering approximation used to derive it, at least in terms of understanding and analysing the information content of the image. The difference between the idealised model given by the first term of equation (1.1) (under the assumption that $\lambda >> L$) and the actual scattering interactions that occur (where $\lambda \sim L$) are compounded in the noise term. The weak scattering approximation essentially allows the scattering model to be linearised which leads to imaging systems being described in terms of a 'linear systems model'. The basis for such models is discussed in the following section.

1.2.1 Linear Systems Modelling

A 'system' may be defined as that which produces a set of output functions from a set of input functions. Physically, it may be an electrical circuit (with input and output voltages, for example) or an imaging system where the inputs and outputs are either complex amplitudes or intensities. From the point of view of 'linear systems theory', the physical nature of the system is unimportant.

Let us represent a system via an operator $\mathcal{L}$ say in terms of the equation

$$s(x,y) = \mathcal{L}[f(x,y)]$$

where f is the input and s is the output. A linear system has the property that

$$\mathcal{L}[af_1(x,y) + bf_2(x,y)] = a\mathcal{L}[f_1(x,y)] + b\mathcal{L}[f_2(x,y)]$$

for all inputs f_1 and f_2 and all constants a and b. Linearity implies that an output function can be broken down into elementary functions, each of which can be separately passed through the system; the total output is then the sum of the 'elementary' outputs.

The 'sampling property' of the delta function allows us to consider any input function to be a linear combination of weighted and displaced delta functions:

$$f(x,y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x',y')\delta(x-x')\delta(y-y')dx'dy'$$

giving an output

$$s(x,y) = \mathcal{L}[f(x,y)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x',y')\mathcal{L}[\delta(x-x')\delta(y-y')]dx'dy'.$$

The system response at (x, y) due to a delta function input at (x', y') is called the Impulse Response Function (IRF) given by

$$p(x, y; x', y') = \mathcal{L}[\delta(x - x')\delta(y - y')].$$

In imaging systems, the quantity p is called the Point Spread Function (PSF). For a linear imaging system,

$$s(x, y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x', y') p(x, y; x', y') dx' dy'$$

If the impulse response function of a linear system depends only on the coordinate differences $(x - x')$ and $(y - y')$, and not on each coordinate separately, i.e.

$$p(x, y; x', y') \equiv p(x - x', y - y'),$$

then we obtain an expression for p which involves the convolution relationship

$$s(x, y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x', y') p(x - x', y - y') dx' dy'.$$

This is an example of a stationary linear system. In optical imaging, for example, a stationary optical systems is called 'isoplanatic'. Isoplanacity requires that the PSF is the same for all field angles and implies that the aberrations are independent of field angle. Many optical imaging systems are (to a good approximation) both linear and isoplanatic.

The convolution relationship between input and output suggests using Fourier transforms (FT) which, via the convolution theorem, gives

$$S(k_x, k_y) = F(k_x, k_y) P(k_x, k_y)$$

where

$$P(k_x, k_y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p(x, y) \exp[-i(k_x x + k_y y)] dx dy$$

i.e.

FT of the output = (FT of the input)$\times$(FT of the impulse response function).

The quantity P is called the system Transfer Function (TF). In optical imaging systems, P is called the Optical Transfer Function or OTF. The OTF is just the 2D FT of the PSF. Note that:

- the convolution relationship only applies to linear stationary imaging systems;
- there is no unique TF for an imaging system with field-dependent aberrations (i.e. for the non-stationary case);
- there is no unique TF for an imaging system when an object is illuminated by spatially partially coherent radiation.

1.2.2 Incoherent Imaging

Consider the case where the object plane is illuminated by an incident wave and by perfectly spatially coherent light. Let the complex amplitude immediately after the object be denoted by $U_{\text{in}}(x,y)$ and $U_{\text{out}}(x,y)$ be the complex amplitude at the image plane. Also, let the complex amplitude at (x,y) in the output due to a unit strength point input be $p(x,y;x',y')$. The total amplitude at (x,y) due to all such points in the object plane is then given by

$$U_{\text{out}}(x,y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} U_{\text{in}}(x',y')p(x,y;x',y')dx'dy'.$$

For an isoplanatic optical system, this reduces to

$$U_{\text{out}}(x,y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} U_{\text{in}}(x',y')p(x-x',y-y')dx'dy'.$$

A spatially coherent optical system is linear in the complex amplitude. Let us now consider the case of narrowband light that is not perfectly spatially coherent. The general complex representation of the time-varying scalar field is called the analytic signal $V(\mathbf{r},t)$; it is defined such that

$$\text{Real scalar field} = \Re[V(\mathbf{r},t)].$$

For narrowband light, the analytic signal can be written in terms of a product of a slowly varying function; the time varying complex amplitude $U(\mathbf{r},t)$ times $\exp(-i\omega t)$. Thus,

$$V(\mathbf{r},t) = U(\mathbf{r},t)\exp(-i\omega t).$$

The instantaneous intensity is defined as

$$I(\mathbf{r},t) = \mid U(\mathbf{r},t) \mid^2$$

whereas the time-averaged intensity $\bar{I}(\mathbf{r})$ (i.e. that observed by an optical detector over a period of time T) is given by

$$\bar{I}(\mathbf{r}) = \frac{1}{2T}\int_{-T}^{T} I(\mathbf{r},t)dt.$$

In general, the time-varying complex amplitudes are related by

$$U_{\text{out}}(x,y,t) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} U_{\text{in}}(x',y',t)p(x,y;x'y')dx'dy'.$$

Coherent illumination implies that $U(x,y,t) = U(x,y)$, i.e. the field does not vary in time. For incoherent light, however, the average intensity is given by

$$\bar{I}_{\text{out}}(x,y) = \frac{1}{2T}\int_{-T}^{T} \mid U_{\text{out}}(x,y,t) \mid^2 dt = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(x,y;x',y')p^*(x,y;x',y')$$

$$\times \left[\frac{1}{2T} \int_{-T}^{T} U_{\text{in}}(x', y', t) U_{\text{in}}^*(x'', y'', t) dt \right] dx' dy' dx'' dy''.$$

The term in [ ] is called the mutual intensity of narrow-band light and is given by

$$J_{\text{in}}(x', y'; x'', y'') = \frac{1}{2T} \int_{-T}^{T} U_{\text{in}}(x', y', t) U_{\text{in}}^*(x'', y'', t) dt$$

or

$$J_{\text{in}}(\mathbf{r}', \mathbf{r}'') = \langle U_{\text{in}}(\mathbf{r}', t) U_{\text{in}}^*(\mathbf{r}'', t) \rangle.$$

Incoherent light is defined to be such that

$$J(\mathbf{r}', \mathbf{r}'') = \bar{I}(\mathbf{r}') \delta(\mathbf{r}' - \mathbf{r}'').$$

That is, two neighbouring points $\mathbf{r}'$ and $\mathbf{r}''$ have uncorrelated fields, for any $\mathbf{r}' \neq \mathbf{r}''$. Using the definition for incoherent light above, the expression for $\bar{I}_{\text{out}}$ becomes

$$\bar{I}_{\text{out}}(x, y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p(x, y; x', y') p^*(x, y; x', y')$$

$$\times \bar{I}_{\text{in}}(x', y') \delta(x' - x'') \delta(y' - y'') dx' dy' dx'' dy''$$

or

$$\bar{I}_{\text{out}}(x, y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \mid p(x, y; x', y') \mid^2 \bar{I}_{\text{in}}(x', y') dx' dy'.$$

where the quantity $\mid p(x, y; x', y') \mid^2$ is the intensity Point Spread Function. For an isoplanatic optical system, this result reduces to

$$\bar{I}_{\text{out}}(x, y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \bar{I}_{\text{in}}(x', y') \mid p(x - x', y - y') \mid^2 dx' dy'.$$

where the bar over I is usually omitted when referring to the intensity because a time average is always assumed.

For perfectly incoherent illumination, an optical system is linear in intensity and, if isoplanicity holds, the output (image) intensity is equal to the input (object) intensity convolved with the intensity point spread function.

1.2.3 Coherent Image Formation

With coherent wavefield, the complex amplitude of the image is equal to that at the object plane convolved with the amplitude point spread function (for an isoplanatic system), i.e.

$$U_{\text{out}}(x, y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} U_{\text{in}}(x', y') p(x - x', y - y') dx' dy'$$

TABLE 1 Schematic diagram illustrating the principles of modelling an imaging system by deriving the imaging equation from the field equations: From the field equations we derive an inhomogeneous wave equation. Using the Green's function together with appropriate boundary condition, we derive an integral equation. From this integral equation, given the geometry of the imaging system and certain approximations (primarily the weak scattering approximation) we derive the imaging equation with expressions for the PSF and the object function in terms of the system parameters and field variables, respectively.

$$\begin{array}{ccccc} & & \text{Maxwell's equation} & & \\ & & \downarrow & & \\ \text{Boundary Conditions} & \rightarrow & \text{Wave equation} & \leftarrow & \text{Green's function} \\ & & \downarrow & & \\ \text{Geometry} & \rightarrow & \text{Integral equation} & \leftarrow & \text{Approximation} \\ & & \downarrow & & \\ & & \text{Imaging equation} & & \end{array}$$

where

$$p(x,y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} P(x',x') \exp\left[-\frac{ik}{z}(xx'+yy')\right] dx'dy'$$

and P is the pupil function of the optical system, i.e. the complex amplitude in the exit pupil. The pupil function P is, for a clear pupil, defined by

$$P(k_x,k_y) = \begin{cases} \exp[ikW(k_x,k_y)], & (k_x,k_y) \in \text{aperture;} \\ 0, & \text{otherwise} \end{cases}$$

where the function W is called the Wave Aberration Function. A shaded or apodized pupil can be handled by introducing an absorption term A,

$$P(k_x,k_y) = A(k_x,k_y)\exp[ikW(k_x,k_y)].$$

Taking the Fourier transform of U_{out} and using the convolution theorem we can write

$$\widetilde{U}_{\text{out}}(k_x,k_y) = \widetilde{U}_{\text{in}}(k_x,k_y)T(k_x,k_y)$$

where $\widetilde{U}_{\text{out}}$ is the spectrum of the image amplitude, $\widetilde{U}_{\text{in}}$ is the spectrum of object amplitude and T is the Coherent Transfer Function.

1.3 EM Imaging Systems Modelling

In terms of developing an EM imaging systems model based on a weak scattering model, Table 1 provides a schematic overview of the interconnecting steps.

The convolution process is fundamental to both general methods of processing a digital image but also in terms of the physical models we use to describe the way in

which images are formed. By studying the 'physics' of an imaging system and using appropriate approximations and geometries, it is possible to formulate the first term in equation (1.1) for which there are two specific classes of images:

- coherent images where

$$I(x,y) =\mid p(x,y) \otimes_2 f(x,y) + n(x,y) \mid^2$$

- incoherent images where

$$I(x,y) =\mid p(x,y) \mid^2 \otimes_2 \mid f(x,y) \mid^2 + \mid n(x,y) \mid^2$$

In the former case the phases associated with the functions f, p and n (all of which may be complex) are mixed via the convolution process and additive operations which is not the case with the generation of an incoherent image.

The overview compounded in Table 1 provides a mathematical description of the PSF and the object function in terms of fundamental physical parameters, which is required in order to understand the information that an image conveys and hence the most appropriate processing methods that should be applied. The mathematical apparatus required for undertaking this task is the basis for Chapters 2-4 which discuss the field equation and wave equation used to model electromagnetic imaging systems (chapter 2), the Green's functions used to solve an inhomogeneous wave equation (Chapter 3) and the analytical method used to compute the scattered field (Chapter 4). Chapter 5 then presents a case study which focuses on modelling a coherent microwave imaging system known as Synthetic Aperture Radar. This case study demonstrates how to combine EM scattering theory with linear systems modelling of an image to generate an expression for the object function of a Synthetic Aperture Radar image in terms of the relative dielectric and conductivity of the scatterer. The key to this approach is to use the weak scattering approximation, thereby linearising the scattering model.

1.4 About this Thesis

This thesis is composed of ten chapters. Chapter 1 provides an introduction to the work presented and sets the context under which the research has been carried out. This chapter specifies the structure of the thesis and some of the original contributions that have been made.

Chapter 2 considers the electromagnetic Langevin-type wave equations required to investigate the scattering of electromagnetic fields from inhomogeneous materials consisting of variations in the relative permittivity, the relative permeability and conductivity. The wave equations derived represent the founding models required to develop a scattering theory that is compatible with Maxwell's macroscopic equation. For this reason, Chapter 2 investigates the properties of EM waves in homogenous and inhomogeneous materials. In the former case, solutions to the

homogenous Helmholtz equation are considered based on the angular spectrum representation of plane waves. This analysis provides a complete representation for the propagation of the scalar electric field potential and the magnetic vector potential under the Lorentz gauge condition.

Chapter 3 provides a background to the working tool of scattering theory, namely, the Green's function. The calculation of Green's function in one-, two- and three-dimensions are considered for the wave equation, the diffusion equation, and finally the Laplace and Poisson equations. The properties of the Green's functions provide an understanding that is necessary to appreciate some of the basic analysis methods required to develop a scattering model especially with regard to the geometry of a scatter imaging system.

Chapter 4 introduces formal scattering theory and looks at the approximations used to develop different scattering models and the analysis of different scattering regimes. The chapter then explores the issues associated with the inverse scattering problem, presenting formal methods to this problem under the weak scattering, weak gradient approximation and (conditional) multiple scattering conditions. An exact inverse scattering approach is then considered which forms the background to material presented in Chapter 6.

Chapter 5 is a case study used to focus the theoretical ideas presented in Chapter 4. This chapter explores the use of weak scattering theory to develop a full model for a Synthetic Aperture Radar (SAR) image which incorporates polarization. The model developed provides a solution to the inverse scattering problem in which SAR images on the different dielectric properties of the ground surface can be obtained by using electric fields with different polarizations. It is also shown how this model can be used to explain the 'sea spikes' phenomenon.

In Chapter 6, the exact inverse scattering solutions developed at the end of Chapter 4 are investigated further and numerical simulations presented to illustrate the superiority of the approach over the weak scattering approximation. These results are then used to develop an expression for a side-band signal for strong (multiple) scattering which is compounded in a single addition term representative of the noise term under the weak scattering approximation. Applications are considered based on SAR data which links to the material discussed in Chapter 5.

Chapter 7 provides an introduction to the theory of EM scattering from random media. Two models are developed:

- weak scattering from random media where the media is taken to conform to a random variable with a defined statistic;

- strong and coherent scattering where the scattered field itself is modelled in terms of a random process leading to a expression for the K-distribution.

However, the principal focus of this chapter is to investigate the use of diffusion based models for multiple scattering and the inverse solutions that are available and applicable in this case.

The diffusion models considered in Chapter 7 provide solutions to the multiple scattering problem when the scattering processes are strong and taken to be

based on application of the diffusion equation. If weak scattering is considered to be a solution to the wave equation (under a the weak scattering approximation), the question arises as to how intermediate scattering can be modelled. Chapter 8 addresses this question by considering a fractional diffusive model. The chapter revisits random walk theory in order to provide a physically intuitive overview of the principles associated with fractional diffusion. Solutions are considered based on the application of fractional calculus and applications studies for the case of light scattering from a tenuous random medium.

The large majority of EM scattering theory assumes that the wavelength is small. Chapter 9 looks at the case when the wavelength of a scalar wavefield as defined by the inhomogeneous Helmholtz equation tends to infinity and it is shown how this condition provides an exact scattering solution. However, the principal purpose of this chapter is to consider the effect of waves scattering from waves over a large frequency spectrum. This leads to the hypothesis that EM and gravitational fields may be unified by attempting to develop a unified wavefield theory rather than unified field theory. The results presented in this chapter lead to a range of arguments, one of which appears to explain why Einstein rings observed in the visible spectrum are blue. This phenomenon is taken to be due to the ability of a gravitational field to not only bend light but to diffract light, a result that has and analogy to the refraction and diffraction of light by a lens, for example.

Chapter 10 of this thesis provides a discussion and conclusion to the work. For completeness, the thesis includes appendices which provide the original research plan, theorems and proofs associated with an exact inverse scattering theory, proof of the relationship between different fractal parameters, a brief introduction to the fractional calculus used in Chapter 8 and finally, a list of open problems associated with the research undertaken to date.

1.5 Original Contributions

The principal and original contributions provided in this thesis are as follows:

- Development of an EM scattering model for Synthetic Aperture Radar based on the weak scattering approximation given in Chapter 5 and an explanation of the sea skies phenomenon.
- Development of an exact inverse scattering solution given towards the end of Chapter 4 and developed further in Chapter 6.
- Modelling intermediate scattering processes in a tenuous random medium using a fractional diffusion model which is the subject of Chapter 8
- Low frequency scattering theory developed in Chapter 9 and the hypotheses and results presented therein.

This material has been published in a range of International Journals and Conferences during the compilation and completion of this thesis. These publication

include the following:

J M Blackledge, *Application of the Fractal Market Hypothesis for Modelling Macroeconomic Time Series*, ISAST, Transactions on Electronics and Signal Processing, **1**(2), 89-110, 2008, ISSN: 1797-2329;
http://www.isastorganization.org/ISAST_ES_1_2008.pdf

J. M. Blackledge, T. Hämäläinen and J. Joutsensalo, *Inverse Scattering Solutions with Applications to Electromagnetic Signal Processing*, ISAST, Transactions on Electronics and Signal Processing (To be published), 2009;
http://eleceng.dit.ie/papers/113.pdf.

J. M. Blackledge, *Scattering from a Tenuous Random Medium with Applications in Optics*, ISAST Transactions on Electronics and Signal Processing (accepted for publication), 2009;
http://eleceng.dit.ie/papers/114.pdf

J. M. Blackledge, *Imaging Reconstruction for Light Scattering from a Tenuous Random Medium*, ISSC 2009, UCD, June 10-11th, 2009;
http://eleceng.dit.ie/papers/115.pdf

J. M. Blackledge, T. Hämäläinen and J. Joutsensalo, *Inverse Scattering Solutions for Side-Band Signals*, ISSC 2009, UCD, June 10-11th, 2009;
http://eleceng.dit.ie/papers/118.pdf

J. M. Blackledge, *Diffusion and Fractional Diffusion Based Image Processing* EGUK Theory and Practice of Computer Graphics, Conference Proceedings, Pages 233 - 240, 2009;
http://eleceng.dit.ie/papers/118.pdf

J. M. Blackledge, *Exact Inverse Schrödinger Scattering*, 19^{th} International Conference on Ion Beam Analysis, Cambridge University, Sept. 7-11, 2009, Preprint Submitted to the IoP Journal of Nuclear Physics B;
http://eleceng.dit.ie/papers/139.pdf

2 ELECTROMAGNETIC FIELDS AND WAVES

This Chapter is concerned with the fundamental equations used to describe the fields that are measured in electromagnetic information (signal and imaging) systems and their relationship with the material variables with which these fields interact. The field equations determine the physical characteristics and behaviour of a particular type of field and the primary purpose of this Chapter is to introduce and discuss the electromagnetic field equations which are employed in later chapters. From these results, we derive wave equations which describe the propagation of electromagnetic waves through homogeneous and inhomogeneous media.

2.1 The Langevin Equation

The propagation of a wavefield can be modelled by various different wave equations depending upon the type of field, the supporting material and its physical state. In general, however, if the supporting material is assumed to be a linear medium and the scalar field $U(\mathbf{r}, t)$ obeys a partial integro-differential equation of the form

$$\hat{D}^{(1)} U(\mathbf{r}, t) = -s(\mathbf{r}, t)$$

where

$$\hat{D}^{(1)} = \hat{D}^{(0)} + \hat{L}$$

and s is a source function. For a vector field $\mathbf{U}(\mathbf{r}, t)$,

$$\hat{D}^{(1)} \mathbf{U}(\mathbf{r}, t) = -\mathbf{s}(\mathbf{r}, t)$$

This is the Langevin equation [1] where $\hat{D}^{(0)}$ and $\hat{L}$ are linear operators: $\hat{D}^{(0)}$ is associated with the homogeneous portion of the medium, $\hat{L}$ is, in general, an integro-differential scattering operator. The source function s describes the emission of the incident field from a given source. The operator $\hat{L}$ models the interaction of the incident field with the differential or local scattering from material inhomogeneities.

The operator $\hat{D}^{(1)}$ has the general form

$$\hat{D}^{(1)} \equiv \hat{D}^{(1)}\left(\nabla, \nabla^2, ...; 1, \frac{\partial}{\partial t}, \frac{\partial^2}{\partial t^2}, ...\right).$$

The source function s is, in general, given by

$$s = p \otimes_{\mathbf{r}} \otimes_t f$$

where f is the probe field, p is a filter weighting function for the emitted field and $\otimes_{\mathbf{r}} \otimes_t$ is taken to denote the convolution over three-dimensional space $\mathbf{r}$ and time t. From this general formalism, one can consider a variety of scalar wave equations governing the propagation of different types of wavefields supported by different isotropic media. Two cases arise that are based on: (i) a rigorous derivation of the Langevin from a set of fundamental field equations; (ii) the proposition of a Langevin equation (a phenomenological model) whose characteristics are confirmed experimentally.

The wave equation that is used to model a wavefield determines the underlying physical model. This includes aspects such as the wavelength or bandwidth of the wavefield, the accuracy of the spatial mapping of the scatter generating parameters in an image (e.g. the level of distortion as determined by the propagation model) and image fuzziness. A fuzzy image is an image which, although attempting to display a specific scatter generating parameter, fails to achieve this because the scattered wavefield that has been measured and processed is corrupted by some other interaction that has not been included on the original model (i.e. the wave equation). Thus, all scatter-imaging techniques are highly model dependent since the reconstruction algorithm is determined by the wave equation which characterizes the medium, in particular, the model associated with the operator $\hat{L}$. An inappropriate choice of wave equation results in image fuzziness. Scatter imaging demands appropriate modelling of the scattering dynamics, even if the computations are approximate. An inexact model will lead to a fuzzy image whereas an approximate computation may lead to poor resolution. Distortion, due to poor propagator models which are compounded in the operator $\hat{D}^{(0)}$ and its associated (free-space) Green function, is a common artifact in many imaging systems and poor physical modelling manifests some form of distortion in most imaging methods. A general criticism, common to many imaging systems, is that emphasis is often placed on a significant amount of computation for image reconstruction and processing. This can provide good, or at least, enhanced resolution but at the expense of developing accurate models for the propagation of a wavefield through the medium that generates the scattered field from which an image is generated and interpreted. This leads to images which are well resolved but may be badly distorted and fuzzy.

In this chapter, we start by considering Maxwell's equations [2] which provide a unified theoretical framework for the interaction of electromagnetic waves with matter. In the case of the macroscopic form of these equations, we introduce material parameters such as the permittivity, the permeability and the conductivity.

The field equations presented provide the fundamental basis for modelling electromagnetic scattering. It is shown how the field equations can (under certain

conditions and approximations) be decoupled to provide a governing inhomogeneous wave equation whose complexity increases according to the number of material parameters that are considered. General methods of solving such equations are then addressed in the following Chapter.

2.2 Maxwell's Equations

We shall now consider Maxwell's equations and study the electromagnetic wavefields and wave equations that arise from these equations [3] We first consider these equations in their microscopic form (for individual charged particles) and go on to consider the macroscopic form of Maxwell's equations (for the case when there are many particles per cubic wavelength) and briefly study the propagation of monochromatic electromagnetic waves in homogeneous media. The macroscopic form of Maxwell's equations is then used to construct inhomogeneous wave equations in a form that are suitable for applying the solutions methods discussed in Chapter 5.

The motions of electrons (and other charged particles) give rise to electric and magnetic fields. These fields are described by the following equations which are a complete mathematical descriptions for the physical laws quoted [3], [4] [1].

Coulomb's law

$$\nabla \cdot \mathbf{e} = 4\pi\rho \tag{2.1}$$

Faraday's law of induction

$$\nabla \times \mathbf{e} = -\frac{1}{c}\frac{\partial \mathbf{b}}{\partial t} \tag{2.2}$$

No free magnetic monopoles exist

$$\nabla \cdot \mathbf{b} = 0 \tag{2.3}$$

Modified (by Maxwell) Ampere's law

$$\nabla \times \mathbf{b} = \frac{1}{c}\frac{\partial \mathbf{e}}{\partial t} + \frac{4\pi}{c}\mathbf{j} \tag{2.4}$$

where $\mathbf{e}$ is the electric field, $\mathbf{b}$ is the magnetic field, $\mathbf{j}$ is the current density, ρ is the charge density and $c \simeq 3 \times 10^8 \ \ \mathrm{ms}^{-1}$ is the speed of light. These microscopic Maxwell's equations are used to predict the pointwise electric $\mathbf{e}$ and magnetic $\mathbf{b}$ fields given the charge and current densities (ρ and $\mathbf{j}$ respectively).

By including a modification to Ampere's law, i.e. the inclusion of the 'displacement current' term $\partial\mathbf{e}/\partial(ct)$, Maxwell provided a unification of electricity and magnetism compounded in the equations above.

[1] For CGS units.

2.2.1 Linearity of Maxwell's Equations

Maxwell's equations are linear because if

$$\rho_1, \quad \mathbf{j}_1 \rightarrow \mathbf{e}_1, \quad \mathbf{b}_1$$

and

$$\rho_2, \quad \mathbf{j}_2 \rightarrow \mathbf{e}_2, \quad \mathbf{b}_2$$

then

$$\rho_1 + \rho_2, \quad \mathbf{j}_1 + \mathbf{j}_2 \rightarrow \mathbf{e}_1 + \mathbf{e}_2, \quad \mathbf{b}_1 + \mathbf{b}_2$$

where $\rightarrow$ means 'produces'. This is because the operators $\nabla\cdot$, $\nabla\times$ and the time derivatives are all linear operators.

2.2.2 Free Space Solution

The solution to these equations is based on exploiting the properties of vector calculus and, in particular, identities involving the curl.

Taking the curl of equation (2.2), we have

$$\nabla \times \nabla \times \mathbf{e} = -\frac{1}{c}\nabla \times \frac{\partial \mathbf{b}}{\partial t}$$

and using the identity

$$\nabla \times \nabla \times \mathbf{e} = \nabla(\nabla \cdot \mathbf{e}) - \nabla^2 \mathbf{e}$$

together with equations (2.1) and (2.4), we get

$$\nabla(4\pi\rho) - \nabla^2 \mathbf{e} = -\frac{1}{c}\frac{\partial}{\partial t}\left(\frac{1}{c}\frac{\partial \mathbf{e}}{\partial t} + \frac{4\pi}{c}\mathbf{j}\right)$$

or, after rearranging,

$$\nabla^2 \mathbf{e} - \frac{1}{c^2}\frac{\partial^2 \mathbf{e}}{\partial t^2} = 4\pi\nabla\rho + \frac{4\pi}{c^2}\frac{\partial \mathbf{j}}{\partial t}. \tag{2.5}$$

Taking the curl of equation (2.4), using the identity above, equations (2.2) and (2.3) and rearranging the result gives

$$\nabla^2 \mathbf{b} - \frac{1}{c^2}\frac{\partial^2 \mathbf{b}}{\partial t^2} = -\frac{4\pi}{c}\nabla \times \mathbf{j}. \tag{2.6}$$

Equations (2.5) and (2.6) are inhomogeneous wave equations for $\mathbf{e}$ and $\mathbf{b}$. These equations are related or coupled to the vector field $\mathbf{j}$ (which is related to $\mathbf{b}$). If we define a region of free space where $\rho = 0$ and $\mathbf{j} = 0$, then both $\mathbf{e}$ and $\mathbf{b}$ satisfy the equation

$$\nabla^2 \mathbf{f} - \frac{1}{c^2}\frac{\partial^2 \mathbf{f}}{\partial t^2} = 0.$$

This is the homogeneous wave equation. One possible solution of this equation (in Cartesian coordinates) is

$$f_x = p(z - ct); \quad f_y = 0, \quad f_z = 0$$

which describes a wave or distribution p moving along z at velocity c. Thus, we have shown that in free space when

$$\nabla \cdot \mathbf{e} = 0, \quad \nabla \cdot \mathbf{b} = 0,$$

$$\nabla \times \mathbf{e} = -\frac{1}{c}\frac{\partial \mathbf{b}}{\partial t}, \quad \nabla \times \mathbf{b} = \frac{1}{c}\frac{\partial \mathbf{e}}{\partial t}.$$

Maxwell's equations describe the propagation of an electric and magnetic (or electromagnetic field) in terms of a wave traveling at the speed of light (see front cover). After developing the origins of the vector calculus, Maxwell derived the wave equations for an electromagnetic field in a paper entitled *A Dynamical Theory of the Electromagnetic Field*, first published in 1865 and arguably one of the greatest intellectual achievements in the history of physics.

2.3 General Solution using the Lorentz Gauge Condition

The solution to Maxwell's equation in free space is specific to the charge density and current density being zero. We now investigate a method of solution for the general case [3], [4], [5]. The basic method of solving Maxwell's equations (i.e. finding $\mathbf{e}$ and $\mathbf{b}$ given ρ and $\mathbf{j}$) involves the following:

- Expressing $\mathbf{e}$ and $\mathbf{b}$ in terms of two other fields ϕ and $\mathbf{A}$.
- Obtaining two separate equations for ϕ and $\mathbf{A}$.
- Solving these equations for ϕ and $\mathbf{A}$ from which $\mathbf{e}$ and $\mathbf{b}$ can then be computed.

For any vector field $\mathbf{A}$

$$\nabla \cdot \nabla \times \mathbf{A} = 0.$$

Hence, if we write

$$\mathbf{b} = \nabla \times \mathbf{A} \tag{2.7}$$

then equation (2.3) remains unchanged. Equation (2.2) can then be written as

$$\nabla \times \mathbf{e} = -\frac{1}{c}\frac{\partial}{\partial t}\nabla \times \mathbf{A}$$

or

$$\nabla \times \left(\mathbf{e} + \frac{1}{c}\frac{\partial \mathbf{A}}{\partial t}\right) = 0.$$

The field $\mathbf{A}$ is called the Magnetic Vector Potential. For any scalar field ϕ

$$\nabla \times \nabla\phi = 0$$

and thus equation (2.2) is satisfied if we write

$$\pm\nabla\phi = \mathbf{e} + \frac{1}{c}\frac{\partial \mathbf{A}}{\partial t}$$

or

$$\mathbf{e} = -\nabla\phi - \frac{1}{c}\frac{\partial \mathbf{A}}{\partial t} \tag{2.8}$$

where the minus sign is taken by convention. ϕ is called the Electric Scalar Potential.

Substituting equation (2.8) into Maxwell's equation (2.1) gives

$$\nabla \cdot \left(\nabla\phi + \frac{1}{c}\frac{\partial \mathbf{A}}{\partial t} \right) = -4\pi\rho$$

or

$$\nabla^2\phi + \frac{1}{c}\frac{\partial}{\partial t}\nabla \cdot \mathbf{A} = -4\pi\rho. \tag{2.9}$$

Substituting equations (2.7) and (2.8) into Maxwell's equation (2.4) gives

$$\nabla \times \nabla \times \mathbf{A} + \frac{1}{c}\frac{\partial}{\partial t}\left(\nabla\phi + \frac{1}{c}\frac{\partial \mathbf{A}}{\partial t} \right) = \frac{4\pi}{c}\mathbf{j}$$

Finally, using the identity

$$\nabla \times \nabla \times \mathbf{A} = \nabla(\nabla \cdot \mathbf{A}) - \nabla^2\mathbf{A}$$

we can write

$$\nabla^2\mathbf{A} - \frac{1}{c^2}\frac{\partial^2 \mathbf{A}}{\partial t^2} - \nabla\left(\nabla \cdot \mathbf{A} + \frac{1}{c}\frac{\partial \phi}{\partial t} \right) = -\frac{4\pi}{c}\mathbf{j} \tag{2.10}$$

2.3.1 Lorentz Gauge Transformation

If we could solve equations (2.9) and (2.10) above for ϕ and $\mathbf{A}$ then $\mathbf{e}$ and $\mathbf{b}$ could be computed. The problem here, is that equations (2.9) and (2.10) are coupled. They can be decoupled by applying a technique known as a 'gauge transformation' called the Lorentz gauge transformation, after Lorentz who was among the first to consider it as an approach to solving these equations. The idea is based on noting that equations (2.7) and (2.8) are unchanged if we let

$$\mathbf{A} \rightarrow \mathbf{A} + \nabla X$$

and

$$\phi \rightarrow \phi - \frac{1}{c}\frac{\partial X}{\partial t}$$

since $\nabla \times \nabla X = 0$. If this gauge function X is taken to satisfy the homogeneous wave equation

$$\nabla^2 X - \frac{1}{c^2}\frac{\partial^2 X}{\partial t^2} = 0$$

then

$$\nabla \cdot \mathbf{A} + \frac{1}{c}\frac{\partial \phi}{\partial t} = 0 \tag{2.11}$$

which is called the Lorentz condition. With equation (2.11), equations (2.9) and (2.10) become

$$\nabla^2\phi - \frac{1}{c^2}\frac{\partial^2 \phi}{\partial t^2} = -4\pi\rho \tag{2.12}$$

and

$$\nabla^2\mathbf{A} - \frac{1}{c^2}\frac{\partial^2 \mathbf{A}}{\partial t^2} = -\frac{4\pi}{c}\mathbf{j}$$

respectively. These equations are non-coupled inhomogeneous wave equations.

2.3.2 Green's Function Solutions

The Green's function solutions to wave equations for ϕ and $\mathbf{A}$ at $(\mathbf{r}_0, t_0)$ (the 'retarded potentials') are given by (see Chapter 3)

$$\phi(\mathbf{r}_0, t_0) = \frac{1}{4\pi\epsilon_0} \int \frac{\rho(\mathbf{r}, \tau)}{\mid \mathbf{r} - \mathbf{r}_0 \mid} d^3\mathbf{r}, \quad \tau = t_0 - \mid \mathbf{r} - \mathbf{r}_0 \mid / c_0$$

and

$$\mathbf{A}(\mathbf{r}_0, t_0) = \frac{\mu_0}{4\pi} \int \frac{\mathbf{j}(\mathbf{r}, \tau)}{\mid \mathbf{r} - \mathbf{r}_0 \mid} d^3\mathbf{r}$$

respectively. These solutions are based on an infinite domain solution in which the boundary conditions associated with spatial support of ϕ and $\mathbf{j}$ are zero (the Neumann boundary condition as discussed in Chapter 3). These Green's function solutions show that a change in ρ and $\mathbf{j}$ affects ϕ and $\mathbf{A}$ $\mid \mathbf{r} - \mathbf{r}_0 \mid / c_0$ seconds later. The change propagates away from the sources ρ and $\mathbf{j}$ at a velocity c_0 which is the theoretical basis for the propagation of electromagnetic waves [6]. Thus, an oscillating point charge in which the charge density can be described by $\rho(\mathbf{r}, t) = \delta^3(\mathbf{r}) \exp(i\omega t)$ where ω is the angular frequency and δ^3 is the three-dimensional delta function, generates an electric field potential given by

$$\phi(r_0, t_0) = \frac{1}{4\pi\epsilon_0 r_0} \exp[i\omega(t_0 - r_0/c_0)], \quad r_0 \equiv \mid \mathbf{r}_0 \mid$$

2.4 Free Space Propagation of EM Waves

Given that an EM wave has been generated by an oscillating point charge, what are the properties of the plane wave propagation (with a directional bias, i.e. propagating in the z direction, for example) in free space when $\rho = 0$ and $\mathbf{j} = \mathbf{0}$? An answer to this question is based on representing plane wave in terms of an 'angular spectrum'.

2.4.1 The Angular Spectrum of Plane Waves

The angular spectrum of planes waves is a way of representing a wavefield in a region of free space. In this representation, any wavefield can be described by a sum (integral) of plane waves travelling in different directions where each plane wave is an elementary solution of the (homogeneous) Helmholtz equation.

$$\left(\nabla^2 - \frac{1}{c^2}\frac{\partial^2}{\partial t^2}\right) U(\mathbf{r}, t) = 0$$

Here, U is taken to represent the electric field potential ϕ or any component of the magnetic vector potential A_x, A_y and A_z.

Consider a scalar monochromatic field

$$u(\mathbf{r}, \omega) = U(\mathbf{r}) \exp(-i\omega t); \quad \mathbf{r} = \hat{\mathbf{x}}x + \hat{\mathbf{y}}y + \hat{\mathbf{z}}z$$

in a source free region $0 \leq z \leq Z$. In the free space domain, the complex amplitude u satisfies the homogeneous Helmholtz equation

$$(\nabla^2 + k^2)u(\mathbf{r}, k) = 0; \;\; k = \frac{\omega}{c}.$$

Let u have the following Fourier representation with respect to (x, y)

$$u(x, y, z) = \mathcal{F}_2[\widetilde{u}] = \int_{-\infty}^{\infty} \widetilde{u}(k_x, k_y, z) \exp[i(k_x x + k_y y)] dk_x dk_y.$$

where $\mathcal{F}_2$ denotes the two-dimensional Fourier transform. Using

$$\mathcal{F}_2[\nabla^2 U] = -(k_x^2 + k_y^2)\widetilde{u} + \frac{\partial^2 \widetilde{u}}{\partial z^2}$$

and substituting into the Helmholtz equation, we obtain

$$-(k_x^2 + k_y^2)\widetilde{u} + \frac{\partial^2 \widetilde{u}}{\partial z^2} + k^2 \widetilde{u} = 0$$

or

$$\frac{\partial^2 \widetilde{u}}{\partial z^2} = -k^2 \widetilde{u} + (k_x^2 + k_y^2)\widetilde{u} = -k_z^2 \widetilde{u}$$

where

$$k_z^2 = k^2 - k_x^2 - k_y^2$$

or

$$k_z = \begin{cases} \sqrt{k^2 - k_x^2 - k_y^2}, & k_x^2 + k_y^2 \leq k^2; \\ i\sqrt{k_x^2 + k_y^2 - k^2}, & k_x^2 + k_y^2 > k^2. \end{cases}$$

The general solution to this equation is of the form

$$\widetilde{u}(k_x, k_y, z) = A(k_x, k_y) \exp(iwz) + B(k_x, k_y) \exp(-iwz)$$

where A and B are arbitrary functions (excluding the degenerate case when $w = 0$). The solution for U can now be written as

$$u(x, y, z) = \int_{-\infty}^{\infty} A(k_x, k_y) \exp[i(k_x x + k_y y + k_z z)] dk_x dk_y$$

$$\int_{-\infty}^{\infty} B(k_x, k_y) \exp[i(k_x x + k_y y - k_z z)] dk_x dk_y.$$

This is the angular spectrum representation of the field. The field U can be considered to be a linear combination of functions

$$\exp[i(k_x x + k_y y \pm k_z z)].$$

Each term in this solution for U represents a plane wave which is a solution of the same differential equation as the field itself (i.e. the Helmholtz equation). Each

term is a mode of the Helmholtz equation and the angular representation is a mode expansion of the Helmholtz equation. The angular spectrum representation is not a Fourier representation, i.e. it is not a 3D Fourier transform of U, rather, it is a superposition of elementary solutions - plane waves - of the Helmholtz equation.

There are four different types of wave:

(i)

$$A(k_x, k_y)\exp[i(k_x x + k_y y + k_z z)];\;\; w = \sqrt{k^2 - k_x^2 - k_y^2},\;\; k_x^2 + k_y^2 \leq k^2$$

where k_z is real, positive or zero (a homogeneous wave propagating from $z = 0$ toward $z = Z$).

(ii)

$$A(k_x, k_y)\exp[i(k_x x + k_y y + k_z z)];\;\; k_x^2 + k_y^2 > k^2$$

where k_z is purely imaginary, i.e.

$$A(k_x, k_y)\exp[i(k_x x + k_y y)]\exp(- \mid k_z \mid z)$$

which describes an inhomogeneous or 'evanescent' wave.

(iii)

$$B(k_x, k_y)\exp[i(k_x x + k_y y - k_z z);\;\; k_x^2 + k_y^2 \leq k^2$$

which describes a homogeneous wave propagating from Z to the origin.

(iv)

$$B(k_x, k_y)\exp[i(k_x x + k_y y - k_z z)];\;\; k_x^2 + k_y^2 > k^2$$

which describes an inhomogeneous wave propagating from Z to 0.

All four of these types of wave are, in general, necessary to represent the field U.

2.4.2 The Half-Space Problem

Consider a half-space where a wave originating from $z = 0$ travels into the space for which $z > 0$ and has an outgoing behaviour. Since the waves are outgoing $B(k_x, k_y) = 0$ and thus in the half plane

$$(x, y, z) = \int_{-\infty}^{\infty} A(k_x, k_y)\exp[i(k_x x + k_y y + k_z z)]dk_x dk_y.$$

If we now let $k_x = kp,\;\; k_y = kq$ and $k_z = km$ where k is the wave number $(= \omega/c)$ then we can write

$$u(x, y, z) = \int_{-\infty}^{\infty} a(p, q)\exp[ik(px + qy + mz)]dpdq$$

where

$$a(p,q) = k^2 A(kp,kq)$$

and

$$m = \begin{cases} \sqrt{1-p^2-q^2}, & p^2+q^2 \le 1; \\ \sqrt{p^2+q^2-1}, & p^2+q^2 > 1. \end{cases}$$

When $p^2+q^2 \le 1$, the mode is a plane wave propagating in a direction whose 'direction cosines' are (p,q,m). Let us consider the relationship between $u(x,y,z)$ and this field at $z=0$. Decomposing the boundary wave (i.e. the wave at $z=0$) into a 2D Fourier integral,

$$u(x,y,0) = \int_{-\infty}^{\infty} \widetilde{u}_0(k_x,k_y) \exp[i(k_x x + k_y y)] dk_x dk_y.$$

According to the angular spectrum representation

$$u(x,y,0) = \int_{-\infty}^{\infty} a(p,q) \exp[ik(px+qy)] dp dq.$$

Substituting $k_x = kp$ and $k_y = kq$,

$$u(x,y,0) = \frac{1}{k^2} \int_{-\infty}^{\infty} a\left(\frac{k_x}{k}, \frac{k_y}{k}\right) \exp[i(k_x x + k_y y)] dk_x dk_y.$$

Thus,

$$\widetilde{u}_0(k_x,k_y) = \frac{1}{k^2} a\left(\frac{k_x}{k}, \frac{k_y}{k}\right)$$

or

$$a(p,q) = k^2 \widetilde{u}_0(kp,kq).$$

Hence,

$$u(x,y,z) = k^2 \int_{-\infty}^{\infty} \widetilde{u}_0(kp,kq) \exp[ik(px+qy+mz)] dp dq$$

$$= \int_{-\infty}^{\infty} \widetilde{u}_0(k_x,k_y) \exp[i(k_x x + k_y y + k_z z) dk_z dk_y.$$

Thus, the spectral amplitude $a(p,q)$ of each plane wave is completely specified by a single spatial frequency component of the boundary value of the field in the plane $z=0$. The frequency of the spatial frequency components are

$$k_x = kp, \quad k_y = kq.$$

Homogeneous waves exist if

$$p^2+q^2 \le 1.$$

Hence, spatial frequencies in the boundary wave such that

$$\left(\frac{k_x}{k}\right)^2 + \left(\frac{k_y}{k}\right)^2 \leq 1$$

or

$$k_x^2 + k_y^2 \leq k^2$$

give rise to homogeneous waves. A spatial frequency k_x arises from a sinusoidal component of period $2\pi/\Delta x$ in the boundary wave. Therefore, periods in the boundary wave such that

$$\frac{1}{(\Delta x)^2} + \frac{1}{(\Delta y)^2} \leq \frac{1}{\lambda^2}$$

give rise to homogeneous waves. Periods such that

$$\frac{1}{(\Delta x)^2} + \frac{1}{(\Delta y)^2} > \frac{1}{\lambda^2}$$

give rise to evanescent waves. Since evanescent waves decay exponentially with distance it follows that detail in $u(x, y, 0)$ smaller than a wavelength is inaccessible in the far field.

2.4.3 The Paraxial Wave Equation

A plane wave propagating along z can be represented by the field

$$u(x, y, z) = A \exp(ikz), \quad k = \frac{\omega}{c}$$

and is unidirectional. An optical beam, taken to be propagating in the z-direction, can be represented by the field

$$u(x, y, z) = \psi(x, y, z) \exp(ikz)$$

where it is assumed that: (i) $\psi(x, y, z)$ varies slowly in comparison with $\exp(ikz)$; (ii) $\psi(x, y, z)$ is concentrated mainly around the axis $(x, y) = (0, 0)$. With these assumptions, an approximate partial differential equation for ψ can be obtained called the Paraxial Wave Equation, whose solution provides a mathematical description for the propagation of an optical (laser) beam.

The field U satisfies the Helmholtz equation

$$\nabla^2 U + k^2 U = 0.$$

Now

$$\frac{\partial^2 U}{\partial x^2} = \frac{\partial^2 \psi}{\partial x^2} \exp(ikz), \quad \text{and} \quad \frac{\partial^2 U}{\partial y^2} = \frac{\partial^2 \psi}{\partial y^2} \exp(ikz).$$

Also,

$$\frac{\partial U}{\partial z} = \frac{\partial \psi}{\partial z} \exp(ikz) + ik\psi \exp(ikz)$$

and

$$\frac{\partial^2 U}{\partial z^2} = \frac{\partial^2 \psi}{\partial z^2}\exp(ikz) + \frac{\partial \psi}{\partial z} ik \exp(ikz)$$
$$+ ik\frac{\partial \psi}{\partial z}\exp(ikz) + (ik)^2 \psi \exp(ikz)$$
$$= \exp(ikz)\left(\frac{\partial^2 \psi}{\partial z^2} + 2ik\frac{\partial \psi}{\partial z} - k^2\psi\right).$$

Assume that ψ varies so slowly with z that

$$\left|\frac{\partial^2 \psi}{\partial z^2}\right| << 2k\left|\frac{\partial \psi}{\partial z}\right|.$$

Under this condition

$$\frac{\partial^2 U}{\partial z^2} \simeq \exp(ikz)\left(2ik\frac{\partial \psi}{\partial z} - k^2\psi\right)$$

and the Helmholtz equation reduces to

$$\frac{\partial^2 \psi}{\partial x^2} + \frac{\partial^2 \psi}{\partial y^2} + 2ik\frac{\partial \psi}{\partial z} = 0.$$

This equation is the paraxial wave equation or beam equation. It can be written in the form

$$\nabla_\perp^2 \psi + 2ik\frac{\partial \psi}{\partial z} = 0$$

where

$$\nabla_\perp^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}$$

is the transverse Laplacian. The condition

$$\left|\frac{\partial^2 \psi}{\partial z^2}\right| << 2k\left|\frac{\partial \psi}{\partial z}\right|$$

implies that

$$\left|\frac{\partial}{\partial z}\left(\frac{\partial \psi}{\partial z}\right)\right| << \frac{4\pi}{\lambda}\left|\frac{\partial \psi}{\partial z}\right|.$$

For small changes, the change $\mid \Delta(\partial\psi/\partial z) \mid$ in $\mid \partial\psi/\partial z \mid$ is such that

$$\left|\frac{\Delta\left(\frac{\partial \psi}{\partial z}\right)}{\Delta z}\right| << \frac{4\pi}{\lambda}\left|\frac{\partial \psi}{\partial z}\right|.$$

If we take $\Delta z = \lambda$, then

$$\left|\frac{\Delta\left(\frac{\partial \psi}{\partial z}\right)}{\frac{\partial \psi}{\partial z}}\right| << 4\pi.$$

Physically, this condition implies that the change in $\frac{\partial \psi}{\partial z}$ over a distance of the order of a wavelength λ is small compared to $\left|\frac{\partial \psi}{\partial z}\right|$ itself.

2.4.4 Solution to the Paraxial Wave Equation

The paraxial wave equation is given by

$$\frac{\partial^2\psi}{\partial x^2}+\frac{\partial^2\psi}{\partial y^2}+2ik\frac{\partial\psi}{\partial z}=0$$

where

$$u(x,y,z)=\psi(x,y,z)\exp(ikz).$$

Let us employ a Fourier integral representation for ψ, i.e.

$$\psi(x,y,z)=\int_{-\infty}^{\infty}\widetilde{\psi}(k_x,k_y,z)\exp[i(k_xx+k_yy)]dk_xdk_y.$$

Substituting this expression into the paraxial wave equation, we get

$$\int_{-\infty}^{\infty}dk_xdk_y\left(\widetilde{\psi}(k_x,k_y,z)[(ik_x)^2+(ik_y)^2]+2ik\frac{\partial\widetilde{\psi}(k_x,k_y,z)}{\partial z}\right)$$

$$\times\exp[i(k_xx+k_yy)]=0$$

and, since this equality holds for all (x,y), it follows that

$$-(k_x^2+k_y^2)\widetilde{\psi}+2ik\frac{\partial\widetilde{\psi}}{\partial z}=0.$$

Rearranging,

$$\frac{1}{\widetilde{\psi}}\frac{\partial\widetilde{\psi}}{\partial z}=\frac{1}{2ik}(k_x^2+k_y^2)$$

or

$$\frac{d}{dz}\ln\widetilde{\psi}=\frac{1}{2ik}(k_x^2+k_y^2)$$

which has the solution

$$\ln\widetilde{\psi}=\frac{1}{2ik}(k_x^2+k_y^2)z+\text{constant}$$

or

$$\widetilde{\psi}(k_x,k_y,z)=\widetilde{\psi}(k_x,k_y,0)\exp\left[\frac{1}{2ik}(k_x^2+k_y^2)z\right].$$

Substituting this result into the Fourier integral representation for ψ, we obtain a general solution to the paraxial wave equation of the form

$$\psi(x,y,z)=\int_{-\infty}^{\infty}\widetilde{\psi}(k_x,k_y,0)\exp\left[-\frac{i}{2k}(k_x^2+k_y^2)z\right]\exp[i(k_xx+k_yy)]dk_xdk_y.$$

Changing variables to $k_x=kp$ and $k_y=kq$, we can write the solution for the amplitude U as

$$U(x,y,z)=\exp(ikz)\int_{-\infty}^{\infty}a(p,q)\exp[ik(px+qy)]\exp\left[-i\frac{k}{2}(p^2+q^2)z\right]dpdq$$

where

$$a(p,q) = k^2\widetilde{\psi}(k_x,k_y,0).$$

This is the solution to the Helmholtz equation in the 'beam approximation'.

2.4.5 Angular Spectrum Representation

Consider a field $u(x,y,z)$ propagating into the half space $z > 0$ given by the Helmholtz equation

$$\nabla^2 U + k^2 U = 0.$$

We may represent the field as an angular spectrum

$$u(x,y,z) = \int_{-\infty}^{\infty} a(p,q)\exp[ik(px+qy+mz)]dpdq$$

where

$$m = \begin{cases} \sqrt{1-p^2-q^2}, & p^2+q^2 \leq 1; \\ i\sqrt{p^2+q^2-1}, & p^2+q^2 > 1, \end{cases}$$

$$a(p,q) = k^2\widetilde{u}_0(kp,kq)$$

and where $\widetilde{u}_0$ is the 2D spatial Fourier transform of the boundary value of U in the plane $z = 0$, i.e.

$$\widetilde{u}_0(k_x,k_y) = \frac{1}{(2\pi)^2}\int_{-\infty}^{\infty} u(x,y,0)\exp[-i(k_x x + k_y y)]dxdy.$$

For the field U to behave like a beam, we require that $\widetilde{u}_0$ must only contain low frequency components such that

$$k_x^2 + k_y^2 << k^2$$

or

$$p^2 + q^2 << 1.$$

The quantity m must therefore be real and positive; it is given approximately by

$$m = (1-p^2-q^2)^{\frac{1}{2}} \simeq 1 - \frac{1}{2}(p^2+q^2).$$

Under this condition, U is given approximately by

$$u(x,y,z) \simeq \exp(ikz)\int_{-\infty}^{\infty} a(p,q)\exp[ik(px+qy)]\exp\left[-i\frac{k}{2}(p^2+q^2)z\right]dpdq.$$

This is the mathematical expression for a beam, subject to the constraint that $a(p,q)$ is appreciable only for values of p and q such that

$$p^2 + q^2 << 1.$$

Note that the field U is expressed in terms of its Fourier transform in the plane $z = 0$. Also note that this expression is identical to the general solution of the paraxial wave equation. The two approaches are mathematically equivalent.

2.5 The Macroscopic Maxwell's Equations

The microscopic form of Maxwell's equations tells us how individual charged particles and electromagnetic fields interact. When there are many particles per cubic wavelength, the electromagnetic radiation 'sees' only a macroscopic average. The medium is then described by its dielectric parameters: the permittivity ϵ, the magnetic permeability μ and the conductivity σ.

Simple averaging of the quantities over a small volume V, e.g.

$$\mathbf{E}(\mathbf{r},t) = \frac{1}{V}\int_V \mathbf{e}(\mathbf{r}',t)d^3\mathbf{r}', \quad \mathbf{B}(\mathbf{r},t) = \frac{1}{V}\int_V \mathbf{b}(\mathbf{r}',t)d^3\mathbf{r}'$$

leads to the following, but not very useful, macroscopic form of Maxwell's equations:

$$\nabla\cdot\mathbf{E} = 4\pi\rho_{\text{macro}}, \quad \nabla\times\mathbf{E} = -\frac{1}{c}\frac{\partial\mathbf{B}}{\partial t},$$

$$\nabla\cdot\mathbf{B} = 0, \quad \nabla\times\mathbf{B} = \frac{1}{c}\frac{\partial\mathbf{E}}{\partial t} + \frac{4\pi}{c}\mathbf{j}_{\text{macro}}.$$

However, both ρ_{macro} and $\mathbf{j}_{\text{macro}}$ can be split into two terms due to free and bound electrons, i.e. we can write

$$\rho_{\text{macro}} = \rho_{\text{bound}} + \rho_{\text{free}}$$

and

$$\mathbf{j}_{\text{macro}} = \mathbf{j}_{\text{bound}} + \mathbf{j}_{\text{free}}.$$

By *bound*, we mean that the electrons are bound to the nucleus to constitute an atom. If we introduce an electric polarization $\mathbf{P}$ of the medium to represent the average dipole moment per unit volume given by

$$\mathbf{P} = -Ne\mathbf{s}$$

where $\mathbf{s}$ is the average vector between bound electrons and nuclei, e is the charge of an electron and N is the average number of electrons per unit volume, then we can define the charge density of bound electrons as

$$\rho_{\text{bound}} = -\nabla\cdot\mathbf{P}$$

and the current density of bound electrons in the form

$$\mathbf{j}_{\text{bound}} = \frac{\partial\mathbf{P}}{\partial t} + c\nabla\times\mathbf{M}$$

where $\mathbf{M}$ is the magnetization vector. At optical frequencies, $\mathbf{M} = 0$ (in the absence of a strong applied magnetic field). Further, we now define the following:

(i) the displacement vector given by $\mathbf{D} = \mathbf{E} + 4\pi\mathbf{P}$;

(ii) the magnetic field strength given by $\mathbf{H} = \mathbf{B} - 4\pi\mathbf{M}$.

From these definitions, we obtain a useful macroscopic form of Maxwell's equations given by [3], [4]

$$\nabla \cdot \mathbf{D} = 4\pi\rho_{\text{free}}, \quad \nabla \times \mathbf{E} = -\frac{1}{c}\frac{\partial \mathbf{B}}{\partial t}$$

$$\nabla \cdot \mathbf{B} = 0, \quad \nabla \times \mathbf{H} = \frac{1}{c}\frac{\partial \mathbf{D}}{\partial t} + \frac{4\pi}{c}\mathbf{j}_{\text{free}}$$

These equations are valid for media which are: (i) non-isotropic; (ii) inhomogeneous.

2.6 EM Waves in a Homogeneous Medium

Having derived Maxwell's equation in macroscopic form, let us now consider the type of solutions they provide for a specific case. Suppose we illuminate a homogeneous material with monochromatic radiation of angular frequency ω. What are the possible solutions of Maxwell's equations in the material? i.e. what waves exist in the medium?

2.6.1 Linear Medium

Assume that all the macroscopic vectors oscillate sinusoidally at angular frequency ω (this is true, in general, only for high frequency, weak fields). Define vector amplitudes $\mathbf{E}(\mathbf{r},t) = \mathbf{E}(\mathbf{r},\omega)\exp(-i\omega t)$, $\mathbf{B}(\mathbf{r},t) = \mathbf{B}(\mathbf{r},\omega)\exp(-i\omega t)$ and so on[2], so that Maxwell's equations can be written in the form

$$\nabla \cdot \mathbf{D} = 4\pi\rho \tag{2.13}$$

$$\nabla \times \mathbf{E} = \frac{i\omega}{c}\mathbf{B} \tag{2.14}$$

$$\nabla \cdot \mathbf{B} = 0 \tag{2.15}$$

$$\nabla \times \mathbf{H} = -\frac{i\omega}{c}\mathbf{D} + \frac{4\pi}{c}\mathbf{j} \tag{2.16}$$

where ρ and $\mathbf{j}$ are taken to be the free charge density and the free current density, respectively.

Let

$$\mathbf{P} = \chi_e \mathbf{E}, \quad \mathbf{M} = \chi_m \mathbf{H} \text{ and } \mathbf{j} = \sigma \mathbf{E}$$

where χ_e is the electric susceptibility, χ_m is the magnetic susceptibility and σ is the conductivity, each of which may be tensors. Note that, in general, this linearity may not occur and $\mathbf{P}$ could be of the form

$$\mathbf{P} = \chi_e \mathbf{E}(1 + a_1\mathbf{E} + a_2\mathbf{E}^2 + ...).$$

[2] Strictly speaking $\mathbf{E}(\mathbf{r},\omega)$, $\mathbf{B}(\mathbf{r},\omega)$, etc. should be given a different notation but, in the context of the equations that follow, it is implied that all dependent variables are functions of $\mathbf{r}$ and ω and not $\mathbf{r}$ and t.

Here $a_1, a_2, ...$ are constant coefficients which would introduce a nonlinear optical material and nonlinear optical effects for example. Note that the effect of introducing monochromatic radiation (i.e. a wavefield oscillating at one single frequency ω) is to replace the time derivatives in Maxwell's equations with $i\omega$ which significantly helps in the algebra required to derive the solutions that follow.

2.6.2 Isotropic Medium

Let χ_e, χ_m and σ be complex scalars (not tensors) and let us define the following:

(i) the dielectric constant given by $\epsilon = 1 + 4\pi\chi_e$;

(ii) the magnetic permeability given by $\mu = 1 + 4\pi\chi_m$

so that we can write

$$\mathbf{D} = \epsilon\mathbf{E} \tag{2.17}$$

and

$$\mathbf{B} = \mu\mathbf{H}. \tag{2.18}$$

Taking the divergence of equation (2.16) and noting that $\nabla \cdot \nabla \times \mathbf{H} = 0$, we have

$$-i\omega\nabla \cdot \mathbf{D} + 4\pi\nabla \cdot \mathbf{j} = 0.$$

Hence, from equation (2.13) we get

$$\rho = -\frac{i}{\omega}\nabla \cdot \mathbf{j} = -\frac{i}{\omega}\nabla \cdot (\sigma\mathbf{E}). \tag{2.19}$$

Substituting equations (2.17), (2.18) and (2.19) into Maxwell's equation (2.13)-(2.16), we obtain the following time independent equations for the complex vector amplitudes:

$$\nabla \cdot \left(\epsilon + \frac{i4\pi\sigma}{\omega}\right)\mathbf{E} = 0, \quad \nabla \times \mathbf{E} = \frac{i\omega\mu}{c}\mathbf{H},$$

$$\nabla \cdot (\mu\mathbf{H}) = 0, \quad \nabla \times \mathbf{H} = -\frac{i\omega}{c}\left(\epsilon + \frac{i4\pi\sigma}{\omega}\right)\mathbf{E}.$$

These equations apply to a linear, isotropic but inhomogeneous medium, i.e. ϵ, μ and σ may be functions of position. Note that, for any vector $\mathbf{X}$ and scalar a,

$$\nabla \cdot (a\mathbf{X}) = \nabla a \cdot \mathbf{X} + a\nabla \cdot \mathbf{X} \neq a\nabla \cdot \mathbf{X}, \text{ unless } \nabla a = 0.$$

2.6.3 Homogeneous Medium

For a homogeneous medium (where ϵ, μ and σ are constants, the previous set of equations reduces to

$$\nabla \cdot \mathbf{E} = 0, \quad \nabla \times \mathbf{E} = \frac{i\omega\mu}{c}\mathbf{H},$$

$$\nabla \cdot \mathbf{H} = 0, \quad \nabla \times \mathbf{H} = -\frac{i\omega}{c}\left(\epsilon + \frac{i4\pi\sigma}{\omega}\right)\mathbf{E}.$$

2.6.4 Plane Wave Solutions

Let

$$\mathbf{E} = \mathbf{E}_0 \exp(i\mathbf{k}_c \cdot \mathbf{r}) \quad \mathbf{H} = \mathbf{H}_0 \exp(i\mathbf{k}_c \cdot \mathbf{r})$$

where $\mathbf{k}_c$ is the complex wave number. Noting that

$$\nabla \cdot [\mathbf{C} \exp(i\mathbf{k}_c \cdot \mathbf{r})] = i\mathbf{k}_c \cdot \mathbf{C} \exp(i\mathbf{k}_c \cdot \mathbf{r})$$

and

$$\nabla \times [\mathbf{C} \exp(i\mathbf{k}_c \cdot \mathbf{r})] = i\mathbf{k}_c \times \mathbf{C} \exp(i\mathbf{k}_c \cdot \mathbf{r})$$

we obtain

$$\mathbf{k}_c \cdot \mathbf{E}_0 = 0, \quad \mathbf{k}_c \cdot \mathbf{H}_0 = 0, \tag{2.20}$$

$$\mathbf{k}_c \times \mathbf{E}_0 = \frac{\mu\omega}{c}\mathbf{H}_0, \tag{2.21}$$

$$\mathbf{k}_c \times \mathbf{H}_0 = -\frac{\omega}{c}\left(\epsilon + \frac{i4\pi\sigma}{\omega}\right)\mathbf{E}_0. \tag{2.22}$$

Equations (2.20) are referred to as the transversality conditions. Substituting equation (2.21) into equation (2.22) yields

$$\frac{c}{\mu\omega}\mathbf{k}_c \times (\mathbf{k}_c \times \mathbf{E}_0) = -\frac{\omega}{c}\left(\epsilon + \frac{i4\pi\sigma}{\omega}\right)\mathbf{E}_0.$$

Using the identity

$$\mathbf{A} \times (\mathbf{B} \times \mathbf{C}) = (\mathbf{A} \cdot \mathbf{C})\mathbf{B} - (\mathbf{A} \cdot \mathbf{B})\mathbf{C}$$

we can write this result in the form

$$(\mathbf{k}_c \cdot \mathbf{E}_0)\mathbf{k}_c - (\mathbf{k}_c \cdot \mathbf{k}_c)\mathbf{E}_0 = -\frac{\mu\omega^2}{c^2}\left(\epsilon + \frac{i4\pi\sigma}{\omega}\right)\mathbf{E}_0$$

or, since $\mathbf{k}_c \cdot \mathbf{E}_0 = 0$, as

$$\mathbf{k}_c \cdot \mathbf{k}_c = n_c^2 k_0^2 \tag{2.23}$$

where

$$k_0 = \frac{2\pi}{\lambda} = \frac{\omega}{c}$$

and

$$n_c^2 = \epsilon\mu + \frac{i4\pi\mu\sigma}{\omega}.$$

Here, n_c is called the complex refractive index. Let

$$n_c = n + i\kappa$$

and

$$\mathbf{k}_c = \mathbf{k} + i\mathbf{a}$$

where n is the refractive index, κ is the extinction index, $\mathbf{k}$ is the wavenumber and $\mathbf{a}$ is the attenuation vector. Substituting these expressions into equation (2.23) and equating the real and imaginary parts gives

$$k^2 - a^2 = k_0^2(n^2 - \kappa^2) \tag{2.24}$$

and

$$\mathbf{k} \cdot \mathbf{a} = k_0^2 n\kappa. \tag{2.25}$$

Thus, plane wave solutions exist of the form

$$\mathbf{E} = \mathbf{E}_0 \exp(i\mathbf{k}_c \cdot \mathbf{r}); \;\; \mathbf{H} = \mathbf{H}_0 \exp(i\mathbf{k}_c \cdot \mathbf{r})$$

where, from equation (2.23),

$$\mid \mathbf{k}_c \mid = n_c k_0, \mathbf{k}_c \cdot \mathbf{E}_0 = 0$$

and

$$\mathbf{H}_0 = \frac{1}{\mu k_0} \mathbf{k}_c \times \mathbf{E}_0.$$

2.6.5 Non-absorbing Media ($\kappa = 0$)

Equations (2.24) and (2.25) reduce to

$$k^2 - a^2 = n^2 k_0^2 > 0$$

and

$$\mathbf{k} \cdot \mathbf{a} = 0.$$

Two kinds of waves are possible:

(i) Real vector waves where $\mathbf{a} = 0, \;\; \mathbf{k}_c = \mathbf{k}, \;\; \mid \mathbf{k} \mid = k_0 n$ and

$$\mathbf{E}(\mathbf{r}, \omega) = \mathbf{E}_0(\mathbf{r}, \omega) \exp(i\mathbf{k} \cdot \mathbf{r})$$

or

$$\mathbf{E}(\mathbf{r}, t) = \mathbf{E}_0(\mathbf{r}, t) \exp[i(\mathbf{k} \cdot \mathbf{r} - \omega t)].$$

This is like a free space plane wave. The velocity of propagation is $\omega/k = c/n$ and the wavelength, is λ/n. Both amplitude and phase are constant and perpendicular to $\mathbf{k}$, i.e. the wave is homogeneous. Since $\mathbf{k} \cdot \mathbf{E}_0 = 0$, the real and imaginary parts of $\mathbf{E}$ are perpendicular to $\mathbf{k}$. $\mathbf{H}$ is also perpendicular to $\mathbf{k}$ and $\Re[\mathbf{E}]$ is perpendicular to $\Re[\mathbf{H}]$.

(ii) Complex wave vector where $\mathbf{k}$ is perpendicular to $\mathbf{a}$ so that $\mathbf{k} \cdot \mathbf{a} = 0$ and

$$\mathbf{E}(\mathbf{r}, t) = \mathbf{E}_0(\mathbf{r}, t) \exp(-\mathbf{a} \cdot \mathbf{r}) \exp[i(\mathbf{k} \cdot \mathbf{r} - \omega t)]$$

which propagates along $\mathbf{k}$ with velocity $\omega/k < c/n$. The amplitude is constant over planes perpendicular to $\mathbf{a}$ and the phase is constant over planes perpendicular to $\mathbf{k}$ - the wave is homogeneous.

2.6.6 Absorbing Media ($\kappa > 0,\ \mathbf{k} \cdot \mathbf{a} \neq 0$)

(i) Homogeneous wave: $\mathbf{k}$ and $\mathbf{a}$ are in the same direction,

$$\mathbf{k} = n\mathbf{k}_0, \quad \mathbf{a} = \kappa\mathbf{k}_0$$

and

$$\mathbf{E}(\mathbf{r}, t) = \mathbf{E}_0(\mathbf{r}, \omega) \exp(-\kappa\mathbf{k}_0 \cdot \mathbf{r}) \exp[i(n\mathbf{k}_0 \cdot \mathbf{r} - \omega t)].$$

This wave propagates along $\mathbf{k}_0$ at velocity c/n, with wavelength λ_0/n and decreases exponentially along the direction of propagation. Both amplitude and phase are constant and perpendicular to $\mathbf{k}_0$ and both $\mathbf{E}$ and $\mathbf{H}$ are perpendicular to $\mathbf{k}_0$. $\Re[\mathbf{E}]$ is not perpendicular to $\Re[\mathbf{H}]$.

(ii) Inhomogeneous wave: $\mathbf{k}$ and $\mathbf{a}$ are not in the same direction. There is constant phase perpendicular to $\mathbf{k}$ and constant amplitude perpendicular to $\mathbf{a}$. Since $\mathbf{a}$ has a component along $\mathbf{k}$, there is a decrease of amplitude along $\mathbf{k}$.

2.7 EM Waves in an Inhomogeneous Medium

In the previous Section, we considered the EM waves that can occur in a homogeneous material that is linear and isotropic by studying Maxwell's equations for monochromatic propagation. We now turn our attention to developing wave equations for a medium that is linear, isotropic and inhomogeneous using Maxwell's equations in the form[3]

$$\nabla \cdot \epsilon\mathbf{E} = \rho, \tag{2.26}$$

$$\nabla \cdot \mu\mathbf{H} = 0, \tag{2.27}$$

$$\nabla \times \mathbf{E} = -\mu\frac{\partial \mathbf{H}}{\partial t}, \tag{2.28}$$

and

$$\nabla \times \mathbf{H} = \epsilon\frac{\partial \mathbf{E}}{\partial t} + \mathbf{j}. \tag{2.29}$$

where $\mathbf{E}(\mathbf{r}, t)$ is the electric field (volts/metre), $\mathbf{H}(\mathbf{r}, t)$ is the magnetic field (amperes/metre), $\mathbf{j}(\mathbf{r}, t)$ is the current density (amperes/metre^2), $\rho(\mathbf{r}, t)$ is the charge density (charge/metre^2), $\epsilon(\mathbf{r})$ is the permittivity (farads/metre) and $\mu(\mathbf{r})$ is the permeability (henries/metre). The values of ϵ and μ in a vacuum (denoted by ϵ_0 and μ_0, respectively) are $\epsilon_0 = 8.854 \times 10^{-12}$ farads/metre and $\mu_0 = 4\pi \times 10^{-7}$ henries/metre. In electromagnetic imaging problems there are two important physical models to consider, based on whether a material is either conductive or non-conductive.

[3] For SI units.

2.7.1 Conductive Materials

In this case, the medium is assumed to be a good conductor. A current is induced which depends on the magnitude of the electric field and the conductivity σ (siemens/metre) of the material from which the object is composed. The relationship between the electric field and the current density is given by Ohm's law

$$\mathbf{j} = \sigma \mathbf{E} \tag{2.30}$$

A good conductor is one where σ is large. By taking the divergence of equation (2.29) and noting that

$$\nabla \cdot (\nabla \times \mathbf{H}) = 0$$

we obtain (using equation (2.26) for constant ϵ)

$$\frac{\partial \rho}{\partial t} + \frac{\sigma}{\epsilon}\rho = 0.$$

The solution to this equation is

$$\rho(t) = \rho_0 \exp(-\sigma t/\epsilon), \quad \text{where} \quad \rho_0 = \rho(t=0).$$

This solution shows that the charge density decays exponentially with time. Typical values of ϵ are $\sim 10^{-12} - 10^{-10}$ farads/metre. Hence, provided σ is not too small, the dissipation of charge is very rapid. It is therefore physically reasonable to set the charge density to zero and, for problems involving the interaction of electromagnetic waves with good conductors, equation (2.26) can be approximated by

$$\nabla \cdot \epsilon \mathbf{E} = 0 \tag{2.31}$$

and equation (2.29) becomes

$$\nabla \times \mathbf{H} = \epsilon \frac{\partial \mathbf{E}}{\partial t} + \sigma \mathbf{E}.$$

Note that, in imaging problems, the material may not necessarily be conductive throughout but may be a varying dielectric with distributed conductive elements. For example, in imaging the surface of the Earth using microwave radiation (Synthetic Aperture Radar), the electromagnetic scattering model is based on a 'ground truth' that is predominantly a dielectric (dry ground surfaces and dry vegetation for example) with distributed conductors (e.g. metallic objects on a dry ground surface, the sea and to a lesser extent rivers and lakes).

2.7.2 Non-conductive Dielectrics

In this case, it is assumed that the conductivity of the medium is negligible and no current can flow, and hence

$$\mathbf{j} = 0$$

and equation (2.29) is just

$$\nabla \times \mathbf{H} = \epsilon \frac{\partial \mathbf{E}}{\partial t}.$$

Also, if the conductivity is zero then $\rho = \rho_0$ and if $\rho_0 = 0$ then equation (2.26) becomes

$$\nabla \cdot \epsilon \mathbf{E} = 0.$$

The issues of when a material is a conductor or a dielectric is compounded in the relative importance of the terms j and $\epsilon(\partial \mathbf{E}/\partial t)$ in equation (2.29). Let us consider the electric and magnetic fields to be monochromatic waves, so that equation (2.29) becomes (with $\mathbf{j} = \sigma \mathbf{E}$)

$$\nabla \times \mathbf{H}(\mathbf{r}, \omega) = (i\omega\epsilon + \sigma)\mathbf{E}(\mathbf{r}, \omega).$$

The relative importance of the terms on the right hand side of equation (2.29) is then determined by the magnitudes of σ and $\omega\epsilon$. If

$$\frac{\sigma}{\omega\epsilon} >> 1$$

then conduction currents dominate and the medium is a conductor. If

$$\frac{\sigma}{\omega\epsilon} << 1$$

then displacement currents dominate and the material behaves as a dielectric. When

$$\frac{\sigma}{\omega\epsilon} \sim 1$$

the material is a quasi-conductor; some types of semi-conductor fall into this category. Note that the ratio $\sigma/\omega\epsilon$ is frequency dependent and that, consequently, a conductor at one frequency may be a dielectric at another. For example, copper has a conductivity of 5.8×10^7 siemens/metre and $\epsilon \simeq 9 \times 10^{-12}$ farads/metre so that

$$\frac{\sigma}{\omega\epsilon} \sim \frac{10^{18}}{\omega}.$$

Up to a frequency of 10^{16}Hz (the frequency of ultraviolet light) $\sigma/\omega\epsilon >> 1$, and copper is a conductor. At a frequency of 10^{20}Hz (the frequency of X-rays), however, $\sigma/\omega\epsilon << 1$ and copper behaves as a dielectric. This is why X-rays travel distances of many wavelengths in copper. An insulator has a conductivity in the order of 10^{-15} siemens/metre and a permittivity of the order of 10^{-11} farads/metre, which gives

$$\frac{\omega\epsilon}{\sigma} \sim 10^4\omega$$

so the conduction current is negligible at all frequencies.

2.7.3 EM Wave Equation

In many electromagnetic imaging systems, the field that is measured is the electric field. It is therefore appropriate to use a wave equation which describes the behaviour of the electric field. This can be obtained by decoupling Maxwell's equations for the

magnetic field $\mathbf{H}$. Starting with equation (2.28), we divide through by μ and take the curl of the resulting equation. This gives

$$\nabla \times \left(\frac{1}{\mu}\nabla \times \mathbf{E}\right) = -\frac{\partial}{\partial t}\nabla \times \mathbf{H}.$$

By taking the derivative with respect to time t of equation (2.29) and using Ohm's law - equation (2.30) - we obtain

$$\frac{\partial}{\partial t}(\nabla \times \mathbf{H}) = \epsilon\frac{\partial^2 \mathbf{E}}{\partial t^2} + \sigma\frac{\partial \mathbf{E}}{\partial t}.$$

From the previous equation we can then write

$$\nabla \times \left(\frac{1}{\mu}\nabla \times \mathbf{E}\right) = -\epsilon\frac{\partial^2 \mathbf{E}}{\partial t^2} - \sigma\frac{\partial \mathbf{E}}{\partial t} \tag{2.32}$$

Expanding the first term, multiplying through by μ and noting that

$$\mu\nabla\left(\frac{1}{\mu}\right) = -\nabla \ln \mu$$

we get

$$\nabla \times \nabla \times \mathbf{E} + \epsilon\mu\frac{\partial^2 \mathbf{E}}{\partial t^2} + \sigma\mu\frac{\partial \mathbf{E}}{\partial t} = (\nabla \ln \mu) \times \nabla \times \mathbf{E}.$$

Expanding equation (2.31) we have

$$\epsilon\nabla \cdot \mathbf{E} + \mathbf{E} \cdot \nabla\epsilon = 0$$

or

$$\nabla \cdot \mathbf{E} = -\mathbf{E} \cdot \nabla \ln \epsilon.$$

Hence, using the vector identity

$$\nabla \times \nabla \times \mathbf{E} = -\nabla^2 \mathbf{E} + \nabla(\nabla \cdot \mathbf{E})$$

we obtain the following wave equation for the electric field

$$\nabla^2 \mathbf{E} - \epsilon\mu\frac{\partial^2 \mathbf{E}}{\partial t^2} - \sigma\mu\frac{\partial \mathbf{E}}{\partial t} = -\nabla(\mathbf{E} \cdot \nabla \ln \epsilon) - (\nabla \ln \mu) \times \nabla \times \mathbf{E}.$$

This equation is inhomogeneous in ϵ, μ and σ. Solutions to this equation provide information on the behaviour of the electric field in a fluctuating conductive dielectric environment. In electromagnetic imaging problems, interest focuses on the behaviour of the scattered EM wavefield generated by variations in the material parameters ϵ, μ and σ. In this context, ϵ, μ and σ are sometimes referred to as the electromagnetic scatter generating parameters. In electromagnetic imaging, the problem is to reconstruct these parameters by measuring certain properties of the scattered electric field. This is a three parameter inverse problem which requires us to first solve for the electric field $\mathbf{E}$ given ϵ, μ and σ.

2.7.4 Inhomogeneous EM Wave Equations

In order to solve the wave equation derived in the last section using the most appropriate analytical methods for imaging science (i.e. Green function solutions which are discussed in the following Chapter), it must be re-cast in the form of the Langevin equation

$$(\nabla^2 + k^2)\mathbf{E} = -\hat{L}\mathbf{E}$$

where $\hat{L}$ is an inhomogeneous differential operator. Starting with equation (2.32), by adding

$$\epsilon_0 \frac{\partial^2 \mathbf{E}}{\partial t^2} - \frac{1}{\mu_0}\nabla \times \nabla \times \mathbf{E}$$

to both sides of this equation and re-arranging, we can write

$$\nabla \times \nabla \times \mathbf{E} + \epsilon_0\mu_0 \frac{\partial^2 \mathbf{E}}{\partial t^2} = -\epsilon_0\mu_0\gamma_\epsilon \frac{\partial^2 \mathbf{E}}{\partial t^2} - \mu_0\sigma\frac{\partial \mathbf{E}}{\partial t} + \nabla \times (\gamma_\mu \nabla \times \mathbf{E})$$

where

$$\gamma_\epsilon = \frac{\epsilon - \epsilon_0}{\epsilon_0} \text{ and } \gamma_\mu = \frac{\mu - \mu_0}{\mu}.$$

We can then use the result (valid for $\rho \sim 0$)

$$\nabla \times \nabla \times \mathbf{E} = -\nabla^2\mathbf{E} + \nabla(\nabla \cdot \mathbf{E})$$

$$= -\nabla^2\mathbf{E} - \nabla(\mathbf{E} \cdot \nabla \ln \epsilon)$$

so that the above wave equation can be written as

$$\nabla^2\mathbf{E} - \epsilon_0\mu_0\frac{\partial \mathbf{E}}{\partial t^2} = \mu_0\epsilon_0\gamma_\epsilon\frac{\partial^2 \mathbf{E}}{\partial t^2} + \mu_0\sigma\frac{\partial \mathbf{E}}{\partial t} - \nabla(\mathbf{E} \cdot \nabla \ln \epsilon) - \nabla \times (\gamma_\mu \nabla \times \mathbf{E}).$$

Finally, introducing the Fourier transform

$$\mathbf{E}(\mathbf{r}, t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \widetilde{\mathbf{E}}(\mathbf{r}, \omega)\exp(i\omega t)d\omega,$$

we can write the above wave equation in the time independent form

$$(\nabla^2 + k^2)\widetilde{\mathbf{E}} = -k^2\gamma_\epsilon\widetilde{\mathbf{E}} + ikz_0\sigma\widetilde{\mathbf{E}} - \nabla(\widetilde{\mathbf{E}} \cdot \nabla \ln \epsilon) - \nabla \times (\gamma_\mu \nabla \times \widetilde{\mathbf{E}})$$

where

$$k = \frac{\omega}{c_0}, \quad c_0 = \frac{1}{\sqrt{\epsilon_0\mu_0}} \text{ and } z_0 = \mu_0 c_0.$$

The parameter z_0 is the free space wave impedance and is approximately equal to 376.6 ohms. The constant c_0 is the velocity at which electromagnetic waves propagate in a perfect vacuum. In electromagnetic imaging, the fundamental problem is to obtain images of the parameters γ_ϵ, γ_μ and the conductivity σ.

2.8 Discussion

This Chapter has been concerned with investigating the field equations for electromagnetic fields. It has been shown how these equations can be reduced or decoupled to provide a linear inhomogeneous scalar wave equation (a Langevin equation) of the form

$$(\nabla^2 + k^2)\mathbf{u}(\mathbf{r}, k) = -\hat{L}\mathbf{u}$$

where $\hat{L}$ is an inhomogeneous linear differential operator which involves the 'scatter generating parameter' sets $(\gamma_\epsilon, \gamma_\mu, \sigma)$ (where the material may be composed of 'good conductors'). In electromagnetism, nonlinear behaviour can occur as a result of the polarization vector having a nonlinear relationship with the electric field vector. The wave equations derived here are the result of trying to find a balance between developing a physical model that is relatively complete but 'simple' enough for the 'forward scattering problem' (solving for the wavefield $\mathbf{u}$ given $\hat{L}$) and the 'inverse scattering problem' (solving for the material parameter sets given $\mathbf{u}$) to become tractable using the analytical methods discussed in the following chapters.

3 GREEN'S FUNCTIONS

Green's functions are named after the mathematician and physicist George Green born in Nottingham in 1793 who 'invented' the Green's function in 1828 [7]. This invention is developed in an essay entitled *Mathematical Analysis to the Theories of Electricity and Magnetism* originally published in Nottingham in 1828 [8] and reprinted by the George Green Memorial Committee to mark the bicentenary of the birth of George Green in 1993 [9].

The Green's function is a powerful mathematical tool rather than a physical concept and was successfully applied to classical electromagnetism and acoustics in the late Nineteenth Century. More recently, the Green's function has been the working tool of calculations in particle physics, condensed matter and solid state physics, quantum mechanics and many other topics of applied mathematics and mathematical physics [10], [11]. Just as the Green's function revolutionized classical field theory in the nineteenth century (hydrodynamics, electrostatics and magnetism) so it revolutionized quantum field theory in the mid-Twentieth Century through the introduction of quantum Green's functions [12]. This provided the essential link between the theories of quantum electrodynamics in the 1940s and 1950s and has played a major role in theoretical physics ever since. For example, Richard Feynman developed the Feynman diagram which was based on the Green's function. In fact, the Feynman diagram can be considered to be a pictorial representation of a Green's function (a Green's function associated with wave operators), referred to by Feynman as a 'propagator'.

Green's functions are used mainly to solve certain types of linear inhomogeneous partial differential equations (although homogeneous partial differential equations can also be solved using this approach). In principle, the Green's function technique can be applied to any linear constant coefficient inhomogeneous partial differential equation (scalar or vector) in any number of independent variables, although in practice difficulties can arise in computing the Green's function analytically. In fact, Green's functions provide more than just a 'solution', for they transform a partial differential equation representation of a physical problem into a integral equation representation of the same problem which is entirely general and complete. The kernel of the integral equation is composed (completely or partly) of the Green's

function associated with the partial differential equation. This is why Green's function solutions are one of the most powerful analytical tools we have for solving partial differential equations, equations that arise in areas of physics such as electromagnetism (Maxwell's equations), wave mechanics (elastic wave equation), optics (Helmholtz equation), quantum mechanics (Schrödinger and Dirac equations), fluid dynamics (fluid equations of motion), relativistic particle dynamics (Klein-Gordon equation) and general relativity (Einstein equations).

With regard to the application of Green's functions, the inverse (scattering) problem is typically based on first defining the (scattering) problem in terms of the solution to a linear inhomogeneous PDE of the type[1]

$$\hat{D}u(\mathbf{r},t) = \hat{L}u(\mathbf{r},t)$$

where u is the wavefield (which is taken to include the scattered wavefield), $\hat{D}$ is a homogeneous differential operator and $\hat{L}$ is an inhomogeneous differential operator. The forward scattering problem can then be defined as follows:

$$\text{Given } \hat{D} \text{ and } \hat{L} \text{ compute } u(\mathbf{r},t).$$

The inverse scattering problem is then defined as:

$$\text{Given } u(\mathbf{r},t) \text{ compute the inhomogeneous characteristics of } \hat{L}.$$

An appropriate Green's function solution (if available) allows us to write (without loss of generality)

$$u(\mathbf{r},t) = \hat{I}\hat{L}u(\mathbf{r},t)$$

where $\hat{I}$ is an integral operator that incorporates the Green's function.

Inverse scattering problems are, in general, concerned with the inversion of integral equations of this type. However, the technique is not limited to wave equations alone, but can be applied to a wide range of inhomogeneous partial differential equations including those associated with diffusion problems (the diffusion equation), transport phenomena (the Föcker-Planck equation) and static problems (e.g. the Poisson equation).

Green's function are relevant to electromagnetic scattering and imaging because electromagnetic waves obey a wave equation. By deriving a Green's function for this equation, we can calculate the response of an EM imaging system at a point to a disturbance at some source. Further, many problems in EM image revolve around the recovery of information that has been degraded or lost in the passage of an EM wave through the medium of propagation. Green's functions give us a very general way of thinking about this process. If we have a good physical model for an imaging process, for example, then we have a better chance of extracting valuable information by reversing the process, this is the essence of the solving the inverse scattering problem using the Green's function solutions.

---

[1] $\mathbf{r}$ is the multi-dimensional space vector and t denotes time.

3.1 Introduction to the Green's Function

In this chapter, we provide an introduction to the Green's function method. The material focuses attention on Green's function solutions to the wave equation. However, for completeness, we also consider Green's function solutions to the diffusion equation and the Poisson and Laplace equations. Here, we are concerned with the use of 'Free Space' Green's functions which provide a general solution in the infinite domain or over a finite domain to which boundary conditions can then be applied.

By way of a short introduction to help the reader understand the principle of using Green's functions we now consider two short examples. The first example is based on considering point sources to generate a solution to an ordinary differential equation and is based on a 'qualitative analysis'. The second example makes specific use of the delta function and its properties to develop a solution which is based on a more systematic analysis - as used throughout this Chapter.

Example 1: Consider the following inhomogeneous ordinary differential equation

$$\hat{D}u(x) = f(x) \tag{3.1}$$

where $\hat{D}$ is a linear differential operator and $f(x)$ is a given function (the source term), the solution being required on the interval $0 \leq x \leq a$ where a is some constant.

Instead of considering $f(x)$ as a continuous source function, let us approximate it by a set of source functions $f(\xi_1), f(\xi_2), ..., f(\xi_n)$ acting at the points $x = \xi_1, x = \xi_2, ..., x = \xi_n$, all for $x \in [0, a]$. Now define the function $g(x, \xi_i)$ to be the solution to equation (3.1) due to a source acting at ξ_i. The solution due to the single effect of this point source is given by $g(x, \xi_i)f(\xi_i)$. The solution for $u(x)$ is then obtained by summing the results for all the n source terms acting over the interval $0 \leq x \leq a$, and takes the form

$$u(x) = \sum_{i=1}^{n} g(x, \xi_i)f(\xi_i).$$

As n becomes larger so that the number of point source functions $f(\xi_i)$ increases, a better and better approximation to $f(x)$ is obtained. In the limit as $n \to \infty$, $|\xi_i - \xi_{i+1}| \to 0 \; \forall i$ and the summation in the equation above may be replaced by an integral to give the required solution to equation (3.1) in the form

$$u(x) = \int_0^a g(x, \xi)f(\xi)d\xi.$$

The function $g(x, \xi)$ is called the Green's function of the problem.

The Green's function is usually denoted by g and G, but the notation changes from author to author. They are usually written in the form $g(x, \xi)$ (as in this example), or $g(\mid x - \xi \mid)$, or $g(x \mid \xi)$ (as used here).

Similar results to the one given above may be obtained for linear partial differential equations. For example, the solution of the Poisson equation in two dimensions, i.e.

$$\nabla^2 u(x,y) = f(x,y); \quad x \in [0,a], \quad y \in [0,b]$$

may be written as

$$u(x,y) = \int_0^a \int_0^b g(x,y;\xi,\eta) f(\xi,\eta) d\xi d\eta$$

where $g(x,y;\xi,\eta)$ is the Green's function of the problem.

The approach to developing a Green's function solution discussed in this example is based on considering point sources to provide a set of elementary results and then summing up the results to give the required solution. In optics, this principle is often referred to as Huygens' principle. It allows the optical field generated by a given source to be computed by considering the field generated from a single point on the source and then summing up the field generated from a large collection of such points. In this sense, the principle behind a Green's function solution is effectively the same as Huygens' principle, i.e. find the solution to the problem for a single point and then integrate over all such points.

A point source can be described by a delta function and the relationship between the delta function and the Green's function is fundamental. By way of a short introduction to the use of the delta function for solving partial differential equations using Green's functions, we consider the following example which, in comparison with the example given above, provides a more complete form of analysis to develop a Green's function solution for the one-dimensional inhomogeneous wave equation.

Example 2: Consider the inhomogeneous wave equation

$$\left(\frac{\partial^2}{\partial x^2} + k^2\right) u(x,k) = f(x) \tag{3.2}$$

where k (the wavenumber) is a constant and $f(x)$ is the source term, the solution being required over all space $x \in (-\infty, \infty)$ subject to the conditions that u and $\partial u/\partial x$ are zero at $\pm\infty$. This equation describes the behaviour of 'steady waves' (with constant wavelength $\lambda = 2\pi/k$) generated by a source $f(x)$. With reference to Example 1, we are considering the case where

$$\hat{D} = \frac{\partial^2}{\partial x^2} + k^2.$$

Let us define the Green's function as being the solution to equation (3.2) when the source term is replaced by a point source or delta function at a point x_0 say, giving the equation

$$\left(\frac{\partial^2}{\partial x^2} + k^2\right) g(x \mid x_0, k) = \delta(x - x_0) \tag{3.3}$$

where δ has the following fundamental property

$$\int_{-\infty}^{\infty} u(x)\delta(x-x_0)dx = u(x_0).$$

Pre-multiplying equation (3.2) by g gives

$$g\left(\frac{\partial^2}{\partial x^2}+k^2\right)u = gf$$

and pre-multiplying equation (3.3) by u gives

$$u\left(\frac{\partial^2}{\partial x^2}+k^2\right)g = u\delta(x-x_0).$$

Subtracting the two results and integrating over all space, we obtain

$$\int_{-\infty}^{\infty}\left(g\frac{\partial^2 u}{\partial x^2}-u\frac{\partial^2 g}{\partial x^2}\right)dx = \int_{-\infty}^{\infty} fgdx - \int_{-\infty}^{\infty} u\delta(x-x_0)dx.$$

Using the generalized sampling property of the delta function given above and rearranging the result, we obtain

$$u(x_0,k) = \int_{-\infty}^{\infty} fgdx - \int_{-\infty}^{\infty}\left(g\frac{\partial^2 u}{\partial x^2}-u\frac{\partial^2 g}{\partial x^2}\right)dx.$$

Evaluating the second integral on the right-hand side,

$$\int_{-\infty}^{\infty}\left(g\frac{\partial^2 u}{\partial x^2}-u\frac{\partial^2 g}{\partial x^2}\right)dx = \int_{-\infty}^{\infty}\left[\frac{\partial}{\partial x}\left(g\frac{\partial u}{\partial x}\right)-\frac{\partial g}{\partial x}\frac{\partial u}{\partial x}-\frac{\partial}{\partial x}\left(u\frac{\partial g}{\partial x}\right)+\frac{\partial u}{\partial x}\frac{\partial g}{\partial x}\right]dx$$

$$= \int_{-\infty}^{\infty}\frac{\partial}{\partial x}\left(g\frac{\partial u}{\partial x}\right)dx - \int_{-\infty}^{\infty}\frac{\partial}{\partial x}\left(u\frac{\partial g}{\partial x}\right)dx = \left[g\frac{\partial u}{\partial x}\right]_{-\infty}^{\infty} - \left[u\frac{\partial g}{\partial x}\right]_{-\infty}^{\infty} = 0$$

provided u and $\partial u/\partial x$ are zero at $x = \pm\infty$. With these conditions, we obtain the Green's function solution in the form

$$u(x_0,k) = \int_{-\infty}^{\infty} f(x)g(x \mid x_0,k)dx.$$

Physically the Green's function associated with wavefield problems, as in this example, represents the way in which a wave propagates from one point in space to another. For this reason, Green's functions are sometimes referred to as propagators. In this case, the Green's function is a function of the 'path length' between x and x_0, irrespective of whether $x > x_0$ or $x < x_0$. The path length is given by $\mid x - x_0 \mid$ and the Green's function is a function of this path length which is why,

using the notation $x \mid x_0 \equiv \mid x - x_0 \mid$, we write $g(x \mid x_0)$. Note that the solution for u is a convolution (since $x \mid x_0 = x_0 \mid x$), the convolution of the source function $f(x)$ with the Green's function $g(\mid x \mid)$. In general, we can consider the solution to an equation of the type

$$\hat{D}u(x) = f(x)$$

to be given by

$$u(x) = g(\mid x \mid) \otimes f(x)$$

where $\otimes$ denotes the convolution operation and g is the solution to the equation

$$\hat{D}u(x) = \delta(x - x_0).$$

Such a solution is of little value unless the Green's function can be computed and, in the following Section, this problem is addressed.

3.2 The Time Independent Wave Operator

In this Section, we shall concentrate on the computation of Green's functions for the time-independent wave equation in one-, two- and three-dimensions. The solution is over all space and the Green's function is not constrained to any particular boundary conditions (except those at $\pm\infty$). It is therefore referred to as a free space Green's function. Green's functions of this type are used in a wide range of physical problems related to the propagation and interaction of waves with matter. They are one of the most important functions in mathematical physics because of the way they allow partial differential equations that describe the interaction of wavefields with matter to be solved. Physically, these Green's functions represent the way in which a wave propagates from one point source to another.

The type of equations that we are forced to consider with regard to the 'physics' of imaging systems, and the analytical techniques that have been developed to cope with them, nearly always originate in some way from the properties of the Green's function that is used. A good understanding of these functions is therefore required if the basic elements of imaging theory are to be understood.

3.2.1 The One-dimensional Green's Function

We start by reconsidering Example 2 given in the Section A1.2 which, through the application of the sampling property of the delta function together with some relatively simple analysis, demonstrated that the solution to the inhomogeneous wave equation

$$\left(\frac{\partial^2}{\partial x^2} + k^2\right) u(x,k) = f(x)$$

for constant k and $x \in (-\infty, \infty)$ subject to the boundary conditions

$$u(x,k) \mid_{\pm\infty} = 0 \quad \text{and} \quad \left[\frac{\partial}{\partial x} u(x,k)\right]_{\pm\infty} = 0$$

is given by

$$u(x_0,k) = \int_{-\infty}^{\infty} f(x)g(x \mid x_0,k)dx$$

where g is the Green's function given by the solution to the equation

$$\left(\frac{\partial^2}{\partial x^2} + k^2\right) g(x \mid x_0,k) = -\delta(x - x_0) \tag{3.4}$$

subject to $g(x|x_0,k)|_{\pm\infty} = 0$ and $[\partial g(x|x_0,k)/\partial x]_{\pm\infty} = 0$. We shall now discuss the evaluation of the Green's function for this case. Note that, here, the Green's function is defined for $-\delta$ on the right hand side instead of δ as used previously. This is for convenience only in the computations that follow; it does not affect the analysis but does reduces the number of negative signs that accompany the calculation. For this reason, many authors define the Green's function with $-\delta$, a definition which is used throughout the rest of this Chapter.

The solution for the Green's function is based on employing the properties of the Fourier transform. Writing $X =\mid x - x_0 \mid$, we express g and δ as Fourier transforms, i.e.

$$g(X,k) = \frac{1}{2\pi}\int_{-\infty}^{\infty} G(u,k)\exp(iuX)du \tag{3.5}$$

and

$$\delta(X) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \exp(iuX)du.$$

Substituting these expressions into equation (3.4) and differentiating gives

$$\frac{1}{2\pi}\int_{-\infty}^{\infty}(-u^2+k^2)G(u,k)\exp(iuX)du = -\frac{1}{2\pi}\int_{-\infty}^{\infty}\exp(iuX)du$$

from which it follows that

$$G(u,k) = \frac{1}{u^2-k^2}.$$

Substituting this result back into equation (3.5), we obtain

$$g(X,k) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\frac{\exp(iuX)}{u^2-k^2}du = \frac{1}{2\pi}\int_{-\infty}^{\infty}\frac{\exp(iuX)}{(u-k)(u+k)}du.$$

The problem is therefore reduced to that of evaluating the above integral. This can be done using Cauchy's integral formula,

$$\oint_C f(z)dz = 2\pi i \times (\text{sum of the residues enclosed by } C)$$

where C is the contour defining the path of integration. In order to evaluate the integral explicitly using this formula, we must consider the singular nature or poles

of the integrand at $z = -k$ and $z = k$. For now, let us consider a contour which encloses both poles. The residue at $z = k$ is given by $\exp(ikX)/(2k)$ and at $z = -k$ by $-\exp(-ikX)/(2k)$. Hence the Green's function is given by

$$g(X,k) = 2\pi i \left(\frac{\exp(ikX)}{4\pi k} - \frac{\exp(-ikX)}{4\pi k} \right) = -\frac{\sin(kX)}{k}.$$

This Green's function represents the propagation of waves travelling away from the point disturbance at $x = x_0$, or 'outgoing waves', and also waves traveling toward the point disturbance, or 'incoming waves'. Since x and x_0 are points along a line, we can consider the result to be the sum of waves travelling to the left of $\delta(x - x_0)$ in which $x < x_0$ and to the right of $\delta(x - x_0)$ where $x > x_0$. In most applications it is convenient to consider the Green's function for outgoing or (more rarely) incoming waves, but not both. Here, the Green's function for a incoming waves is given by

$$g(x \mid x_0, k) = -\frac{i}{2k} \exp(-ik \mid x - x_0 \mid)$$

and for outgoing waves is

$$g(x \mid x_0, k) = \frac{i}{2k} \exp(ik \mid x - x_0 \mid).$$

3.2.2 The Two-dimensional Green's Function

A Green's function in two- and three-dimensions is synonymous with the source-observer system illustrated in Figure 1. If the position of the source is denoted by $\mathbf{r}_0$ and the position of the observer by $\mathbf{r}$, then the Green's function is written as a function of $\mid \mathbf{r} - \mathbf{r}_0 \mid$ where in Cartesian coordinates,

$$\mid \mathbf{r} - \mathbf{r}_0 \mid = \sqrt{(x - x_0)^2 + (y - y_0)^2}.$$

When the functional dependence of the Green's function is declared, instead of writing $g(\mid \mathbf{r} - \mathbf{r}_0 \mid)$, which is messy, it is more convenient to write $g(\mathbf{r}, \mathbf{r}_0)$ or $g(\mathbf{r} \mid \mathbf{r}_0)$. Here, the latter notation is used throughout, i.e.

$$g(\mathbf{r} \mid \mathbf{r}_0) \equiv g(\mid \mathbf{r} - \mathbf{r}_0 \mid).$$

In two dimensions, the same method can be used to obtain the (free space) Green's function as that used to solve the one-dimensional case, i.e. to solve the equation

$$(\nabla^2 + k^2) g(\mathbf{r} \mid \mathbf{r}_0, k) = -\delta^2(\mathbf{r} - \mathbf{r}_0)$$

where

$$\mathbf{r} = \hat{\mathbf{x}} x + \hat{\mathbf{y}} y, \quad \mathbf{r}_0 = \hat{\mathbf{x}} x_0 + \hat{\mathbf{y}} y_0,$$

and

$$\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}.$$

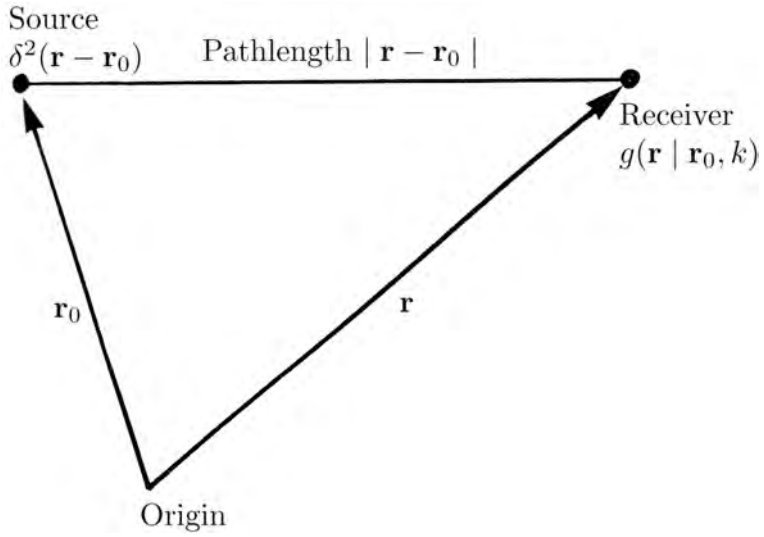

FIGURE 1 Source-observer geometry used to defined the Green's function which is a function of the 'pathlength' $\mid \mathbf{r}-\mathbf{r}_0 \mid$.

Note that

$$\delta^2(\mathbf{r}-\mathbf{r}_0) \equiv \delta(x-x_0)\delta(y-y_0).$$

Also note that g is a function of the path length $\mid \mathbf{r}-\mathbf{r}_0 \mid$. Writing $\mathbf{R} = \mathbf{r}-\mathbf{r}_0$ and using the same technique as before, namely the one used to derive an integral representation of the one-dimensional Green's function, we obtain

$$g(R,k) = \frac{1}{(2\pi)^2} \int_{-\infty}^{\infty} \frac{\exp(i\mathbf{u}\cdot\mathbf{R})}{u^2-k^2} d^2\mathbf{u}.$$

In polar coordinates this result becomes

$$g(R,k) = \frac{1}{(2\pi)^2} \int_0^{\pi} \int_{-\infty}^{\infty} \frac{\exp(iuR\cos\theta)}{u^2-k^2} u\,du\,d\theta.$$

Integrating over u first and using Cauchy's residue theorem, we have

$$\oint_C \frac{z\exp(izR\cos\theta)}{(z+k)(z-k)} dz = i\pi\exp(ikR\cos\theta)$$

where the contour of integration C has been chosen to enclose just one of the poles at $z=k$. This provides an expression for the 'outgoing' Green's function in which the wave propagates away from the point disturbance at $\mathbf{r}_0$. A solution for the pole at $z=-k$ would provide a solution which represents a wavefield converging on $\mathbf{r}_0$. The 'outgoing' Green's function is usually the most physically significant result (except for an implosion for example). Thus, the (outgoing) Green's function can be written in the form

$$g(R,k) = \frac{i}{4\pi} \int_0^{\pi} \exp(ikR\cos\theta) d\theta.$$

Writing the Green's function in this form allows us to employ the result

$$H_0^{(1)}(kR) = \frac{1}{\pi}\int_0^{\pi} \exp(ikR\cos\theta)d\theta$$

where $H_0^{(1)}$ is the Hankel function (of the first kind and of order zero). This is the integral representation for the Hankel transform and it can be used to write the two-dimensional Green's function as

$$g(\mathbf{r} \mid \mathbf{r}_0, k) = \frac{i}{4}H_0^{(1)}(k \mid \mathbf{r} - \mathbf{r}_0 \mid).$$

Unlike the one-dimensional (Section 3.2.1) and three-dimensional (Section 3.2.3) Green's functions, the two-dimensional Green's function can not be expressed explicitly in terms of a complex exponential function. For this reason, it is usual to implement a semi-classical limit which is based on the series [13]

$$\frac{i}{4}H_0^{(1)}(kR) =$$

$$\frac{i}{4}\sqrt{\frac{2}{\pi kR}}\left[1 - \frac{8i}{kR} - \frac{9}{128(kR)^2} + ... + \frac{[(2n-1)!!]^2}{(8i)^n n!(kR)^n} + ...\right]\exp[i(kR - \pi/4)]$$

It is common to use only the first term and consider the Green's function to be given by

$$g(R,k) = \frac{i}{4}\sqrt{\frac{2}{\pi kR}}\exp[i(kR - \pi/4)]$$

or

$$g(\mathbf{r} \mid \mathbf{r}_0, k) = \frac{1}{\sqrt{8\pi}}\exp(i\pi/4)\frac{\exp(ik \mid \mathbf{r} - \mathbf{r}_0 \mid)}{\sqrt{k \mid \mathbf{r} - \mathbf{r}_0 \mid}}$$

Formally, this expression is based on the condition $kR \rightarrow$ and is an asymptotic approximation to the two-dimensional Green's function This condition means that the wavelength of the wave originating from $\mathbf{r}_0$ is small compared with the distance between $\mathbf{r}_0$ and $\mathbf{r}$, which is physically reasonable with regard to high frequency electromagnetic scattering.

3.2.3 The Three-dimensional Green's Function

In three dimensions, the free space Green's function is given by the solution to the equation

$$(\bigtriangledown^2 + k^2)g(\mathbf{r} \mid \mathbf{r}_0, k) = -\delta^3(\mathbf{r} - \mathbf{r}_0)$$

where

$$\mathbf{r} = \hat{\mathbf{x}}x + \hat{\mathbf{y}}y + \hat{\mathbf{z}}z, \quad \mathbf{r}_0 = \hat{\mathbf{x}}x_0 + \hat{\mathbf{y}}y_0 + \hat{\mathbf{z}}z_0,$$
$$\delta^3(\mathbf{r} - \mathbf{r}_0) \equiv \delta(x - x_0)\delta(y - y_0)\delta(z - z_0)$$

and

$$\bigtriangledown^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}.$$

In this case, following the same procedure as before,

$$g(R,k) = \frac{1}{(2\pi)^3} \int_{-\infty}^{\infty} \frac{\exp(i\mathbf{u}\cdot\mathbf{R})}{u^2-k^2} d^3\mathbf{u}.$$

It proves convenient to evaluate this integral using spherical polar coordinates which gives

$$g(R,k) = \frac{1}{(2\pi)^3} \int_0^{2\pi} d\phi \int_{-1}^{1} d(\cos\theta) \int_0^{\infty} \frac{\exp(iuR\cos\theta)u^2}{u^2-k^2} du.$$

Integrating over ϕ and θ we then obtain

$$g(R,k) = \frac{1}{2\pi^2 R} \int_0^{\infty} \frac{u\sin(uR)}{u^2-k^2} du.$$

Since the integrand is an even function we may extend the integration to include the interval $-\infty$ to 0 by writing

$$g(R,k) = \frac{1}{4\pi^2 R} \int_{-\infty}^{\infty} \frac{u\sin(uR)}{u^2-k^2} du.$$

This is done in anticipation of using Cauchy's residue theorem to evaluate the contour integral

$$\oint_C \frac{z\exp(izR)}{(z-k)(z+k)} dz$$

which has simple poles at $z = \pm k$. Choosing the contour C to enclose the pole at $z = k$ (the 'outgoing' case), the residue is

$$\frac{\exp(ikR)}{2}$$

and, thus, the 'outgoing' Green's becomes

$$g(\mathbf{r} \mid \mathbf{r}_0, k) = \frac{1}{4\pi \mid \mathbf{r}-\mathbf{r}_0 \mid} \exp(ik \mid \mathbf{r}-\mathbf{r}_0 \mid).$$

We see that in one-, two- and three dimensions the Green's function is singular. The precise nature of the singularity changes from one dimension to the next. In three dimensions, the Green's function is spatially singular when $\mathbf{r} = \mathbf{r}_0$, whereas in one dimension the singularity occurs when $k = 0$. In two dimensions, a singularity occurs when either $k = 0$ or $\mathbf{r} = \mathbf{r}_0$. An example of this two-dimensional Green's function is observed when a small stone falls vertically into a large pool of water. The symmetrical expanding wavefront represents the result of applying a short impulse to the surface of the water. What is observed is a good approximation to a Hankel function! There are relatively few examples in nature which are characteristic of an ingoing Green's function since most impulses produce wavefields that propagate

away from the point of disturbance. One notable example of an ingoing Green's function is an implosion.

In addition to the derivation of the 3D Green's function given above, the function can be derived in a more physically intuitive way. Imagine a point source of radiation with a wavenumber k which gives out a stream of waves, moving radially outwards. If the distance from the source is R, then we should expect to be able to describe the wavefield (taken to be of unit amplitude) as $\exp(ikR)$ in the usual way. However, in 3D, the intensity of the field should obey an inverse square law and vary as $1/R^2$. But the intensity of the wavefield is proportional to the modulus squared and so the amplitude of the wavefield must be proportional to $1/R$ and hence the amplitude of the field should be

$$\frac{\exp(ikR)}{R}$$

which is the correct form of the Green's function, up to a numerical factor (i.e. $1/4\pi$). Finally, if the source is taken to be at $\mathbf{r}_0$ and the field is measured at $\mathbf{r}$, then R must be given in terms of the 'path length' $\mid \mathbf{r} - \mathbf{r}_0 \mid$.

3.2.4 Asymptotic Forms

Although the Green's functions for the inhomogeneous wave equation can be computed in the manner already discussed, their algebraic form is not always easy, useful or indeed necessary to work with. This is because the geometry of many imaging systems justifies an approximation. For this reason, it is now appropriate to consider the form of the Green's function when the field generated by a point source is moved away from that source, i.e. when the magnitude of $\mathbf{r}_0$ becomes increasingly larger than the magnitude of $\mathbf{r}$. There are two approximations which are important in this respect which are often referred to as the Fraunhofer and Fresnel approximations. These approximations are usually associated with the applications of Green's functions in optics (in which both Fraunhofer and Fresnel undertook their original work) but are, in fact, of general applicability.

The Fraunhofer Approximation

In one-dimension, we note that

$$|x - x_0| = \begin{cases} x_0 - x, & x_0 > x; \\ x - x_0, & x > x_0. \end{cases}$$

so that the Green's function for a left-travelling wave for example can be written as

$$g(x \mid x_0, k) = \frac{i}{2k} \exp(ikx_0) \exp(-ikx), \quad x_0 > x$$

and

$$g(x \mid x_0, k) = \frac{i}{2k} \exp(-ikx_0) \exp(ikx), \quad x_0 < x$$

for a right-travelling wave.

In two- and three-dimensions, we expand the path length between the source and observer in terms of their respective coordinates. To start with, let us look at the result in two dimensions. In this case,

$$\mid \mathbf{r} - \mathbf{r}_0 \mid = \sqrt{r_0^2 + r^2 - 2\mathbf{r}\cdot\mathbf{r}_0} = r_0 \left(1 - \frac{2\mathbf{r}\cdot\mathbf{r}_0}{r_0^2} + \frac{r^2}{r_0^2}\right)^{\frac{1}{2}}$$

where $\mathbf{r} = \hat{\mathbf{x}}x + \hat{\mathbf{y}}y$, $r = \mid \mathbf{r} \mid$ and $r_0 = \mid \mathbf{r}_0 \mid$. A binomial expansion of this result gives

$$\mid \mathbf{r} - \mathbf{r}_0 \mid = r_0 \left(1 - \frac{\mathbf{r}\cdot\mathbf{r}_0}{r_0^2} + \frac{r^2}{2r_0^2} + ...\right) \tag{3.6}$$

which under the condition

$$\frac{r}{r_0} << 1$$

reduces to

$$\mid \mathbf{r} - \mathbf{r}_0 \mid \simeq r_0 - \hat{\mathbf{n}}_0 \cdot \mathbf{r}$$

where

$$\hat{\mathbf{n}}_0 = \frac{\mathbf{r}_0}{r_0}.$$

It is sufficient to let

$$\frac{1}{\mid \mathbf{r} - \mathbf{r}_0 \mid} \simeq \frac{1}{r_0}, \quad r_0 >> r$$

because small changes in $\hat{\mathbf{n}}\cdot\mathbf{r}$ compared to r_0 are not significant in an expression of this type. However, with the exponential function

$$\exp[ik(r_0 - \hat{\mathbf{n}}_0 \cdot \mathbf{r})]$$

a relatively small change in the value of $r_0 - \hat{\mathbf{n}}_0 \cdot \mathbf{r}$ compared to r_0 will still cause this term to oscillate rapidly, particularly if the value of k is large. We therefore write

$$\exp(ik \mid \mathbf{r} - \mathbf{r}_0 \mid) = \exp(ikr_0)\exp(-ik\hat{\mathbf{n}}_0 \cdot \mathbf{r}).$$

The asymptotic form of the two dimensional Green's function is then given by

$$g(\mathbf{r} \mid \mathbf{r}_0, k) = \frac{\exp(i\pi/4)}{\sqrt{8\pi}} \frac{1}{\sqrt{kr_0}} \exp(ikr_0)\exp(-ik\hat{\mathbf{n}}_0 \cdot \mathbf{r}), \quad kr_0 >> 1.$$

In three dimensions, the result is (using exactly the same arguments as in the two dimensional case)

$$g(\mathbf{r} \mid \mathbf{r}_0, k) = \frac{1}{4\pi r_0} \exp(ikr_0)\exp(-ik\hat{\mathbf{n}}_0 \cdot \mathbf{r})$$

where

$$\mathbf{r} = \hat{\mathbf{x}}x + \hat{\mathbf{y}}y + \hat{\mathbf{z}}z.$$

When we observe the field described by a Green's function at large distances (i.e. the wavefield generated by a point source a long distance away), it behaves

like a plane wave $\exp(-ik\hat{\mathbf{n}}_0 \cdot \mathbf{r})$. Approximating the Green's function in this way provides a description for the wave in what is commonly referred to as the far field or Fraunhofer zone (or plane). This approximation is often referred to as the Fraunhofer approximation in physical optics. In this zone, the wave front which reaches the observer is a plane wave front because, in effect, the divergence of the field is so small. Observations of a field in this zone are said to be in the Fourier plane because they lead to equations that involve a Fourier transform as shall be shown later. This is the basis for Fraunhofer diffraction theory which is important in applications such as X-ray crystallography, electromagnetic and acoustic imaging and, of course, modern optics.

The Fresnel Approximation

When the source is brought closer to the observer, the wavefront ceases to be a plane wavefront. In this case, the Fraunhofer approximation is inadequate and another approximation for the Green's function must be used. This is known as the Fresnel approximation and is based on incorporating the next term in the binomial expansion of $\mid \mathbf{r} - \mathbf{r}_0 \mid$, namely the quadratic term $r^2/2r_0^2$ in equation (3.6). In this case, it is assumed that $r^2/r_0^2 << 1$ rather than $r/r_0 << 1$ so that all the terms in the binomial expansion of $\mid \mathbf{r} - \mathbf{r}_0 \mid$ that occur after the quadratic term can be neglected. As before, $\mid \mathbf{r} - \mathbf{r}_0 \mid^{-1}$ is approximated by $1/r_0$ but the exponential term now possesses an additional feature, namely a 'quadratic phase factor'. In this case, the two and three-dimensional Green's functions are given by

$$g(\mathbf{r} \mid \mathbf{r}_0, k) = \frac{\exp(i\pi/4)}{\sqrt{8\pi}} \frac{\exp(ikr_0)}{\sqrt{kr_0}} \exp(-ik\hat{\mathbf{n}}_0 \cdot \mathbf{r}) \exp(ir^2/2r_0), \quad kr_0 >> 1$$

and

$$g(\mathbf{r} \mid \mathbf{r}_0, k) = \frac{\exp(ikr_0)}{4\pi r_0} \exp(-ik\hat{\mathbf{n}}_0 \cdot \mathbf{r}) \exp(ir^2/2r_0)$$

respectively. This type of approximation is used in the study of systems (optical systems for example) in which the divergence of the field is a measurable quantity. This is important in imaging systems such as Synthetic Aperture Radar, the application of Fresnel- or zone-plates for example, and Fresnel optics in general. If the source is moved even closer to the observer then neither the Fraunhofer nor the Fresnel approximations will apply. In such cases, it is usually easier to retain the Green's function in full rather than consider another term in the binomial expansion of the path length. Analysis of a wavefield that is produced when a non-asymptotic form of the Green's function is used is referred to as near field analysis. Thus, the Green's function solution to two- and three-dimensional wave type partial differential equations usually falls into one of the three categories:

(i) near field analysis;

(ii) intermediate field (Fresnel zone) analysis;

(iii) far field (Fraunhofer zone of Fourier plane) analysis.

These 'zones' are characterized by the geometry of the 'wavefront' as illustrated in Figure 2.

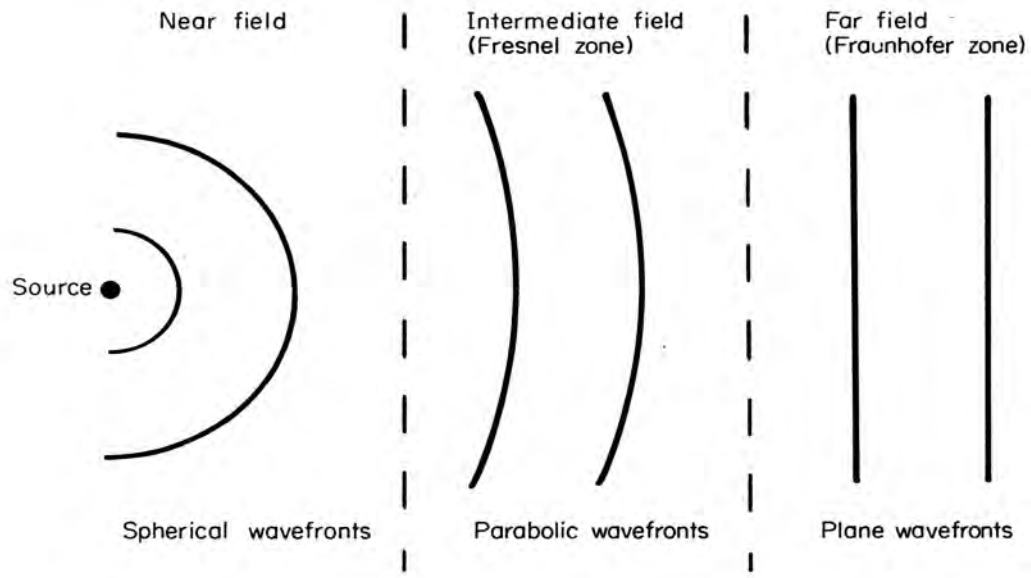

FIGURE 2 Characteristic wavefronts in the near, intermediate and far fields

In practice the far field approximation is much easier to use. This is because it leads to solutions that can be written in terms of a Fourier transform which is a relatively easy transform to work with and invert. Using the Fresnel approximation leads to solutions which involve a class of integral known as the Fresnel integral. The nonlinear behaviour of this integral, because of the quadratic phase factor, makes it more difficult to evaluate compared with the Fourier integral. There are relatively few applications in wavefield theory which require a full near field analysis. This is fortunate, because near field analysis presents some formidable analytical and computational problems.

3.3 Wavefields Generated by Sources

Now that we have studied Green's functions for the inhomogeneous time-independent wave equation, in this Section we turn our attention to the more general problem of developing a solution for the wavefield $u(\mathbf{r},k)$ generated by an arbitrary and time independent source function $f(\mathbf{r})$. This study is a prelude to the work discussed in the following Chapter which provides an introduction to scattering theory in which the source function is not $f(\mathbf{r})$ but $f(\mathbf{r})u(\mathbf{r},k)$ or $k^2 f(\mathbf{r})u(\mathbf{r},k)$. Working in three dimensions, our aim is to solve

$$(\nabla^2 + k^2)u(\mathbf{r},k) = -f(\mathbf{r}), \quad \mathbf{r} \in V$$

for u where V is the volume of the source function which is of compact support (occupies a finite region of space V) . Outside of this region, it is assumed that the

source function is zero. Note that we define the source term as $-f$ rather than $+f$. This is done so that there is consistency with the definition of the Green's function which is defined in terms of $-\delta$ by convention. We start by writing the equation for a Green's function, i.e.

$$(\nabla^2 + k^2)g(\mathbf{r} \mid \mathbf{r}_0, k) = -\delta^3(\mathbf{r} - \mathbf{r}_0).$$

If we now multiply both sides of the first equation by g and both sides of the second equation by u, then by subtracting the two results we obtain

$$g\nabla^2 u - u\nabla^2 g = -gf + u\delta^3.$$

By integrating the last equation over all space, we can exploit the result

$$\int_{-\infty}^{\infty} u(\mathbf{r}, k)\delta^3(\mathbf{r} - \mathbf{r}_0)d^3\mathbf{r} = u(\mathbf{r}_0, k)$$

and therefore write (noting that $\mathbf{r} \in V$)

$$u(\mathbf{r}_0, k) = \int_V f(\mathbf{r})g(\mathbf{r} \mid \mathbf{r}_0, k)d^3\mathbf{r}$$

$$+ \int_V [g(\mathbf{r} \mid \mathbf{r}_0, k)\nabla^2 u(\mathbf{r}, k) - u(\mathbf{r}, k)\nabla^2 g(\mathbf{r} \mid \mathbf{r}_0, k)]d^3\mathbf{r}.$$

We see that this expression is not a 'proper solution' for u because u occurs on both the left and right hand sides. What we require is a solution for u in terms of known quantities on the right hand side of the above equation. To this end, we can simplify the second term by using Green's theorem

$$\int_V (g\nabla^2 u - u\nabla^2 g)d^3\mathbf{r} = \oint_S (g\nabla u - u\nabla g) \cdot \hat{\mathbf{n}} d^2\mathbf{r}.$$

Here, S defines the surface enclosing the volume V and $d^2\mathbf{r}$ is an element of this surface. The unit vector $\hat{\mathbf{n}}$ points out of the surface and is perpendicular to the surface element $d^2\mathbf{r}$. Green's theorem is a special but important consequence of Gauss' divergence theorem as shown below.

3.3.1 Green's Theorem

Let u and g be any two piecewise continuous functions of position and S be a surface enclosing a volume V. If u, g and their first and second partial derivatives are single-valued and continuous within and on S, then

$$\int_V (g\nabla^2 u - u\nabla^2 g)d^3\mathbf{r} = \oint_S \left(g\frac{\partial u}{\partial \hat{\mathbf{n}}} - u\frac{\partial g}{\partial \hat{\mathbf{n}}} \right) d^2\mathbf{r}$$

where $\partial/\partial\hat{\mathbf{n}}$ is a partial derivative in the outward normal direction on S.

The proof of this result stems from noting that since

$$\nabla \cdot (g\nabla u) = \nabla g \cdot \nabla u + g\nabla^2 u$$

and

$$\nabla \cdot (u\nabla g) = \nabla u \cdot \nabla g + u\nabla^2 g$$

then

$$\int_V \nabla \cdot (g\nabla u - u\nabla g) d^3\mathbf{r} = \int_V (g\nabla^2 u - u\nabla^2 g) d^3\mathbf{r}.$$

But from Gauss' theorem

$$\int_V \nabla \cdot \mathbf{F} d^3\mathbf{r} = \oint_S \mathbf{F} \cdot \hat{\mathbf{n}} d^2\mathbf{r}$$

for any vector $\mathbf{F}$. Hence,

$$\int_V \nabla \cdot (g\nabla^2 u - u\nabla^2 g) d^3\mathbf{r} = \oint_S (g\nabla u - u\nabla g) \cdot \hat{\mathbf{n}} d^2\mathbf{r}$$

which provides the basic result, a result that can be written in an alternative (and arguably more elegant way) by defining

$$\nabla u \cdot \hat{\mathbf{n}} \equiv \frac{\partial u}{\partial \hat{\mathbf{n}}}$$

and

$$\nabla g \cdot \hat{\mathbf{n}} \equiv \frac{\partial g}{\partial \hat{\mathbf{n}}}$$

so that we can write

$$\int_V (g\nabla^2 u - u\nabla^2 g) d^3\mathbf{r} = \oint_S \left(g\frac{\partial u}{\partial \hat{\mathbf{n}}} - u\frac{\partial g}{\partial \hat{\mathbf{n}}} \right) d^2\mathbf{r}.$$

This theorem provides a solution for the wavefield u at $\mathbf{r}_0$ of the form

$$u(\mathbf{r}_0, k) = \int_V fg d^3\mathbf{r} + \oint_S (g\nabla u - u\nabla g) \cdot \hat{\mathbf{n}} d^2\mathbf{r}.$$

We have shown that using a Green's function and Green's theorem, the solution to the equation

$$(\nabla^2 + k^2)u(\mathbf{r}, k) = -f(\mathbf{r}), \quad \mathbf{r} \in V$$

is

$$u(\mathbf{r}_0, k) = \oint_S (g\nabla u - u\nabla g) \cdot \hat{\mathbf{n}} d^2\mathbf{r} + \int_V fg d^3\mathbf{r}.$$

It is important to appreciate that this solution is entirely general with no conditions being placed on any of the analysis at any point other than that u are g are piecewise continuous. However, as discussed before, it is not a 'solution' as such because the field variable u occurs on both the left hand and right hand sides of the 'solution'.

It is therefore better to consider this 'solution' to be a transform from a partial differential equation to an integral equation. From a mathematical point of view, a Green's function is that function which allows any linear inhomogeneous PDE to be transformed to an integral equation. Finally, note that the homogeneous equation

$$(\nabla^2 + k^2)u(\mathbf{r}, k) = 0$$

also has a Green's function solution given by

$$u(\mathbf{r}_0, k) = \oint_S (g\nabla u - u\nabla g) \cdot \hat{\mathbf{n}} d^2\mathbf{r}.$$

3.3.2 Dirichlet and Neumann Boundary Conditions

Although Green's theorem allows us to simplify the solution for the wavefield u derived in the previous Section (in the sense that we now have a two dimensional instead of a three dimensional integral), we still do not have a proper solution for u since this field variable is present on both the left and right hand sides of the integral equation for u. However, as a result of applying Green's theorem, we now only need to specify u and ∇u on the surface S. Therefore, if we know, *a priori*, the behaviour of u and ∇u on S, then we can compute u at any other observation point $\mathbf{r}_0$. Clearly, a statement about the nature of u and ∇u on S is required, i.e. the boundary conditions need to be specified.

In general, the type of conditions that may be applied depends on the applications that are involved. In practice, two types of boundary conditions are commonly considered. The first one, known as the homogeneous Dirichlet boundary condition, states that u is zero on S. The second one, known as the homogeneous Neumann condition, states that ∇u is zero on S. Taken together, these boundary conditions are known as the 'homogeneous conditions' and are referred to as such throughout the rest of this work. When u satisfies these homogeneous boundary conditions, the solution for u is given by

$$u(\mathbf{r}_0, k) = \int_V f(\mathbf{r}) g(\mathbf{r} \mid \mathbf{r}_0, k) d^3\mathbf{r}$$

because

$$\oint_S (g\nabla u - u\nabla g) \cdot \hat{\mathbf{n}} d^2\mathbf{r} = 0.$$

If the wavefield generated by a source is measured a long distance away from the location of the source, then by using the far field approximation for the Green's function, we have

$$u(\hat{\mathbf{n}}_0, k) = \frac{1}{4\pi r_0} \exp(ikr_0) \int_V f(\mathbf{r}) \exp(-ik\hat{\mathbf{n}}_0 \cdot \mathbf{r}) d^3\mathbf{r}.$$

In this case, the field generated by the source is given by the 3D Fourier transform of the source function f. By measuring the radiation pattern produced by a source

denoted by $f(\mathbf{r})$, the structure or spatial distribution of the source may be recovered through a processes of inversion. In the far field, the source function can be recovered by taking the inverse Fourier transform of the observed field. This is an example of a solution to a class of problem known as an inverse source problem.

3.3.3 The Reciprocity Theorem

The reciprocity theorem applies to all Green's functions associated with any linear partial differential equation. Here, the theorem will be proved for the 3D Green's function corresponding to the time-independent wave equation. The theorem states that if $\mathbf{r}_1$ and $\mathbf{r}_2$ are two points in space then

$$g(\mathbf{r}_1 \mid \mathbf{r}_2, k) = g(\mathbf{r}_2 \mid \mathbf{r}_1, k).$$

The proof of this result which can be obtained by considering the equations

$$(\nabla^2 + k^2) g(\mathbf{r} \mid \mathbf{r}_1, k) = -\delta^3(\mathbf{r} - \mathbf{r}_1), \quad r \in V$$

and

$$(\nabla^2 + k^2) g(\mathbf{r} \mid \mathbf{r}_2, k) = -\delta^3(\mathbf{r} - \mathbf{r}_2) \quad r \in V.$$

Then

$$\begin{aligned} &g(\mathbf{r} \mid \mathbf{r}_2, k)\nabla^2 g(\mathbf{r} \mid \mathbf{r}_1, k) - g(\mathbf{r} \mid \mathbf{r}_1, k)\nabla^2 g(\mathbf{r} \mid \mathbf{r}_2, k) \\ &= g(\mathbf{r} \mid \mathbf{r}_1, k)\delta^3(\mathbf{r} - \mathbf{r}_2) - g(\mathbf{r} \mid \mathbf{r}_2, k)\delta^3(\mathbf{r} - \mathbf{r}_1) \end{aligned}$$

Integrating over V and using Green's theorem, for homogeneous boundary conditions on the surface of V, we have

$$\int_V g(\mathbf{r} \mid \mathbf{r}_1, k)\delta^3(\mathbf{r} - \mathbf{r}_2) d^3\mathbf{r} - \int_V g(\mathbf{r} \mid \mathbf{r}_2, k)\delta^3(\mathbf{r} - \mathbf{r}_1) d^3\mathbf{r} = 0$$

or

$$g(\mathbf{r}_2 \mid \mathbf{r}_1, k) = g(\mathbf{r}_1 \mid \mathbf{r}_2, k).$$

Thus, the propagation of a wave from a point at $\mathbf{r}_1$ to $\mathbf{r}_2$ is the same as the propagation of a wave from a point at $\mathbf{r}_2$ to $\mathbf{r}_1$.

3.4 Time Dependent Green's Function

We have studied the Green's function for the time independent wave equation. In this section, we investigate the time dependent case.

As an introduction to the time dependent Green's function, let us first consider the case where we have a homogeneous source of scalar radiation a long distance away from an observer at $\mathbf{r}$. Here, the scalar wavefield u as a function of space $\mathbf{r}$ and time t is described by the homogeneous equation

$$\left(\nabla^2 + \frac{1}{c^2}\frac{\partial^2}{\partial t^2}\right) U(\mathbf{r}, t) = 0 \tag{3.7}$$

where c is the velocity at which the radiation propagates from the source to the observer.

3.4.1 Continuous Wave Sources

Let us assume that the source emits a continuous wave which oscillates at a fixed frequency. In this case, the source is known as a continuous wave (CW) or monochromatic source as used earlier in Chapter 4 to investigate the solutions to the Maxwell's equations for a linear isotropic medium. The time dependence of the radiation field is described by the complex exponential function $\exp(i\omega t)$ where ω is the angular frequency ($= 2\pi \times$frequency). The time dependent field u can therefore be written as

$$U(\mathbf{r}, t) = u(\mathbf{r}, \omega) \exp(i\omega t).$$

Substituting this expression into equation (3.7), we obtain

$$(\nabla^2 + k^2)u(\mathbf{r}, k) = 0$$

where

$$k = \frac{\omega}{c} = \frac{2\pi}{\lambda}$$

is the wavenumber and λ is the wavelength of the wavefield described by the function u. A solution to this equation is

$$u(\mathbf{r}, k) = \exp(i\mathbf{k} \cdot \mathbf{r})$$

where the wave vector

$$\mathbf{k} = k\hat{\mathbf{n}}$$

and it is assumed that the amplitude of the wave is 1. The unit vector $\hat{\mathbf{n}}$ points along the direction in which the wave propagates. Thus, the solution for the time dependent wavefield becomes

$$U(\mathbf{r}, t) = \exp[i(\mathbf{k} \cdot \mathbf{r} + \omega t)].$$

However, an equally valid solution is

$$U(\mathbf{r}, t) = \exp[i(\mathbf{k} \cdot \mathbf{r} - \omega t)]$$

which is obtained by using the $\exp(-i\omega t)$ to describe the time dependence of the wavefield. If we imagine a straight line along the direction of $\hat{\mathbf{n}}$, then the above solution for u represents a wave propagating to the right whereas the former solution represents a wave propagating to the left. The function

$$\exp[i(\mathbf{k} \cdot \mathbf{r} + \omega t)]$$

is said to describe a left-travelling wave and

$$\exp[i(\mathbf{k} \cdot \mathbf{r} - \omega t)]$$

is referred to as a right-travelling wave.

3.4.2 Pulsed Sources

If the source emits a pulse of radiation, then the time dependent field can be written as the sum of many different monochromatic waves of different frequencies ω and amplitudes u. If we consider all the different possible frequencies that can exist between $-\infty$ and ∞, then $U(\mathbf{r},t)$ can be written in terms of its Fourier transform as,

$$U(\mathbf{r},t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} u(\mathbf{r},\omega)\exp(i\omega t)d\omega.$$

Here, U describes a left-travelling pulse. We can also consider a solution for a right-travelling pulse by writing

$$U(\mathbf{r},t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} u(\mathbf{r},\omega)\exp(-i\omega t)d\omega.$$

Substituting either of these expressions into equation (3.7), we obtain

$$(\nabla^2 + k^2)u(\mathbf{r},k) = 0$$

where now k is not fixed but can take on any value between $-\infty$ and ∞. The time dependent field produced by a left-travelling pulse is therefore

$$U(\mathbf{r},t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \exp[i(\mathbf{k}\cdot\mathbf{r} + \omega t)]d\omega.$$

If we now write $\mathbf{k}\cdot\mathbf{r}$ as $k\hat{\mathbf{n}}\cdot\mathbf{r} = (\omega/c)\hat{\mathbf{n}}\cdot\mathbf{r}$, then, using the integral representation for a delta function, the above equation can be written as

$$U(\mathbf{r},t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \exp[i\omega(t + \hat{\mathbf{n}}\cdot\mathbf{r}/c)]d\omega = \delta(t + \hat{\mathbf{n}}\cdot\mathbf{r}/c).$$

The expression for a right-travelling pulse is given by

$$U(\mathbf{r},t) = \delta(t - \hat{\mathbf{n}}\cdot\mathbf{r}/c).$$

3.5 Time Dependent Sources

Let us now turn our attention to the case when an inhomogeneous time varying source produces a wavefield $U(\mathbf{r},t)$. To describe this situation mathematically, we introduce a source function $f(\mathbf{r},t)$. The wavefield is then governed by the inhomogeneous equation

$$\left(\nabla^2 + \frac{1}{c^2}\frac{\partial^2}{\partial t^2}\right)U(\mathbf{r},t) = -f(\mathbf{r},t).$$

Now, as in the time dependent case, the Green's function describes the wavefield that is produced when the source function is a delta function, i.e. when

$$S(\mathbf{r}, t) = \delta^n(\mathbf{r} - \mathbf{r}_0)\delta(t - t_0)$$

where $n = 1, 2$ or 3 depending on whether we are considering a one-, two-, or three-dimensional wavefield respectively. Hence, the equation for the time dependent Green's function (which is usually denoted by G) is given by

$$\left(\nabla^2 + \frac{1}{c^2}\frac{\partial^2}{\partial t^2}\right) G(\mathbf{r} \mid \mathbf{r}_0, t \mid t_0) = -\delta^n(\mathbf{r} - \mathbf{r}_0)\delta(t - t_0).$$

3.5.1 3D Time Dependent Green's Function

We shall consider the three-dimensional time-dependent problem first which is based on an evaluation using the time-independent Green's function. We write G and $\delta(t - t_0)$ as Fourier transforms,

$$G(\mathbf{r} \mid \mathbf{r}_0, t \mid t_0) = \frac{1}{2\pi}\int_{-\infty}^{\infty} g(\mathbf{r} \mid \mathbf{r}_0, \omega)\exp[i\omega(t - t_0)]d\omega$$

and

$$\delta(t - t_0) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \exp[i\omega(t - t_0)]d\omega$$

where ω is the angular frequency. Substituting these equations into the equation for G we then have

$$(\nabla^2 + k^2)g(\mathbf{r} \mid \mathbf{r}_0, k) = -\delta^3(\mathbf{r} - \mathbf{r}_0)$$

which is the same equation as that used previously to define the time-independent Green's function. Thus, once g has been obtained, the time dependent Green's function can be derived by computing the Fourier integral given above. Using the expression for g derived earlier,

$$G(\mathbf{r} \mid \mathbf{r}_0, t \mid t_0) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{1}{4\pi \mid \mathbf{r} - \mathbf{r}_0 \mid}\exp(ik \mid \mathbf{r} - \mathbf{r}_0 \mid)\exp[i\omega(t - t_0)]d\omega$$

$$= \frac{1}{4\pi \mid \mathbf{r} - \mathbf{r}_0 \mid}\delta(t - t_0 + \mid \mathbf{r} - \mathbf{r}_0 \mid / c).$$

3.5.2 2D Time Dependent Green's Function

In two dimensions, the point source (which depends on x and y) can be treated as a line source, that is a uniform source extending from $z_0 = -\infty$ to $z_0 = \infty$ along a line parallel to the z axis and passing through the point (x_0, y_0). Thus, a

simple way of computing the two-dimensional Green's function is to integrate the three-dimensional Green's function from $z_0 = -\infty$ to $z_0 = \infty$, i.e.

$$G(\mathbf{s} \mid \mathbf{s}_0, t \mid t_0) = \int_{-\infty}^{\infty} \frac{\delta(t - t_0 + \mid \mathbf{r} - \mathbf{r}_0 \mid /c)}{4\pi \mid \mathbf{r} - \mathbf{r}_0 \mid} dz_0$$

where

$$\mathbf{s} = \hat{\mathbf{x}}x + \hat{\mathbf{y}}y$$

and

$$\mathbf{s}_0 = \hat{\mathbf{x}}x_0 + \hat{\mathbf{y}}y_0.$$

Writing $\tau = (t - t_0)c$, $\xi = z_0 - z$, $S = \mid \mathbf{s} - \mathbf{s}_0 \mid$ and $R = \mid \mathbf{r} - \mathbf{r}_0 \mid$ we have

$$R^2 = \xi^2 + S^2$$

and

$$\frac{dR}{dz_0} = \frac{\xi}{R}$$

and so the Green's function can be written in the form

$$G(S, \tau) = \frac{1}{4\pi} \int_{-\infty}^{\infty} \frac{\delta(\tau + R)}{\sqrt{R^2 - S^2}} dR$$

$$= \begin{cases} \frac{1}{4\pi} \frac{1}{\sqrt{\tau^2 - S^2}}, & \tau > S; \\ 0, & \tau < S. \end{cases}$$

3.5.3 1D Time Dependent Green's Function

In one dimension, the time-dependent Green's function can be calculated by integrating the three dimensional Green's function over z_0 and y_0. Alternatively, we can use the expression for $g(x \mid x_0, k)$ (right-travelling Green's function) giving

$$G(x \mid x_0, t \mid t_0) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{i}{2k} \exp(ik \mid x - x_0 \mid) \exp[i\omega(t - t_0)] d\omega.$$

This equation is the inverse Fourier transform of the product of two functions (given that $k = \omega/c$), namely $i/2k$ and $\exp(ik \mid x - x_0 \mid)$. Thus, using the convolution theorem and noting that

$$\frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{i}{2k} \exp[i\omega(t - t_0)] d\omega = \frac{c}{4} \operatorname{sgn}(t - t_0)$$

and

$$\frac{1}{2\pi} \int_{-\infty}^{\infty} \exp(ik \mid x - x_0 \mid) \exp[i\omega(t - t_0)] d\omega = \delta(t - t_0 + \mid x - x_0 \mid /c),$$

we obtain

$$G(x \mid x_0, t \mid t_0) = \frac{c}{4}\,\mathrm{sgn}(t - t_0) \otimes \delta(t - t_0 + \mid x - x_0 \mid /c)$$

$$= \frac{c}{4}\,\mathrm{sgn}[t - t_0 + \mid x - x_0 \mid /c]$$

where $\otimes$ denotes the convolution integral and sgn is defined by

$$\mathrm{sgn}(x) = \begin{cases} 1, & x > 0; \\ -1, & x < 0. \end{cases}$$

3.5.4 Comparison of the Time-Dependent Green's Functions

There is a striking difference between the time dependent Green's functions derived in the last Sections. In three dimensions, the effect of an impulse after a time $t - t_0$ is found concentrated on a sphere of radius $c(t - t_0)$ whose centre is the source point. The effect of the impulse can therefore only be experienced by an observer at one location over an infinitely short period of time. After the pulse has passed by an observer, the disturbance ceases. In two dimensions, the disturbance is spread over the entire plane $\mid \mathbf{s} - \mathbf{s}_0 \mid$. At $\mid \mathbf{s} - \mathbf{s}_0 \mid = c(t - t_0)$ there is a singularity which defines the position of the two dimensional wavefront as it propagates outwards from the source point at $\mathbf{s}_0$. For $\mid \mathbf{s} - \mathbf{s}_0 \mid < c(t - t_0)$ the Green's function is still finite and therefore, unlike the three-dimensional case, the disturbance is still felt after the wavefront has passed by the observer. In one dimension, the disturbance is uniformly distributed over all points of observation through which the wavefront has passed, since for all values of $\mid x - x_0 \mid$ and $c(t - t_0)$, the Green's function is either $c/4$ or $-c/4$. This is illustrated in Figure 3.

Compared with the Green's function in one and two dimensions, the three dimensional Green's function possesses the strongest singularity. Compared to the delta function, the singularity of the two-dimensional Green's function at $\mid \mathbf{s} - \mathbf{s}_0 \mid = c(t - t_0)$ is very weak. In one dimension, the time dependent Green's function is not singular but discontinuous when $\mid x - x_0 \mid = c(t - t_0)$.

With regard to the two-diemsnional Green's function, the time-dependent form (with $S = R$ and $\tau = t$)

$$G(R, t) = \frac{1}{4\pi} \frac{1}{\sqrt{t^2 - R^2}}$$

has spectral characteristic defined by $iH_0^{(1)}(kR)/4$. The series expression for this Green's function presented in Section 3.2.2 corresponds, term by term, to the following expansion of the time-domain solution (with $\tau \equiv t - R$,

$$\frac{1}{4\pi\sqrt{t^2 - R^2}} = \frac{1}{4\pi\sqrt{\tau(2R + \tau)}} = \frac{1}{4\pi\sqrt{2R}} \tau^{-1/2} \left(1 + \frac{\tau}{2R}\right)^{-\frac{1}{2}}$$

$$= \frac{1}{4\pi\sqrt{2R}} \left(\tau^{-1/2} - \frac{1}{4R}\tau^{1/2} + \ldots + \frac{\Gamma\left(n + \frac{1}{2}\right) \tau^{n-1/2}}{\Gamma\left(\frac{1}{2}\right) \Gamma(n)(-2R)^n} + \ldots \right)$$

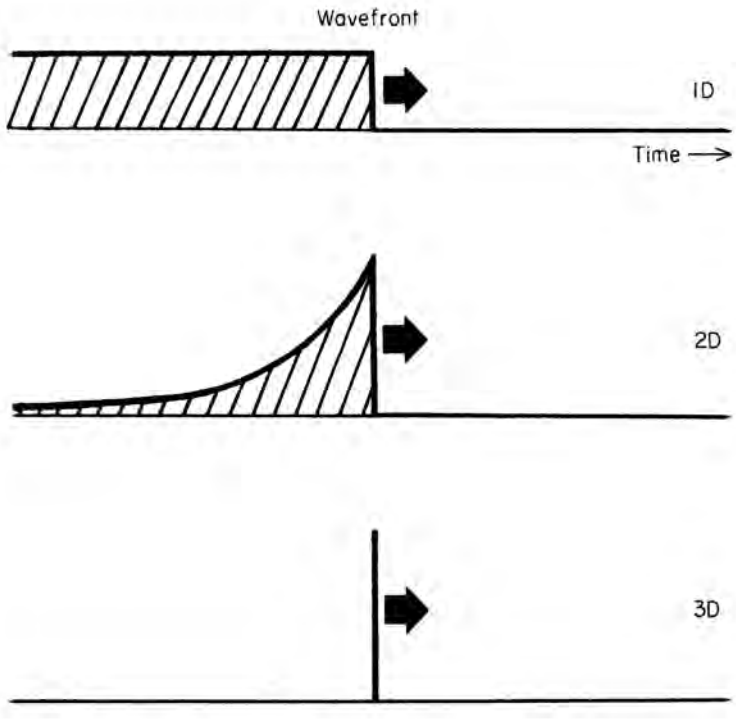

FIGURE 3 Time history of the Green's function in one, two and three dimensions

where we consider only the non-vanishing region $\tau > 0$. This is the Frobenius series (with respect to t) about the singular point $t = r$ (the first arrival time). This is the Taylor series offset by the factor $(t-r)^{-1/2}$ which accounts for the non-analyticity. The correspondence of these series is based on the Fourier transform of the power law $(i\omega)^{-n}, n > -1$ given by

$$\frac{\tau^{n-1}}{\Gamma(n)} = \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{1}{(i\omega)^n}\exp(i\omega\tau)d\omega, \quad \Gamma(n+1) = n!$$

which is based on the Laplace transform of τ^n, i.e.

$$\int_0^{\infty} \tau^n \exp(-p\tau)d\tau = \frac{n!}{p^{n+1}}$$

with inverse Laplace transform

$$\frac{1}{2\pi i}\int_{-i\infty}^{i\infty} \frac{n!}{p^{n+1}}\exp(p\tau)dp = \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{n!}{(i\omega)^{n+1}}\exp(i\omega\tau)d\omega, \quad p = i\omega$$

Thus, the Fourier transform of the n^{th} term of above series is

$$\frac{(-1)^n i^{n+1/2}\Gamma\left(n+\frac{1}{2}\right)^2}{(kR)^{n+1/2}(2\pi)^{n+3/2}n!}$$

which, using the property $\Gamma(n+1/2) = \sqrt{\pi}(2n+1)!!/2^n$ for integer n, matches the n^{th} term (including pre-factors) of the series for $iH_0^{(1)}(kR)/4$. The first term of this series corresponds to assuming a $(t-R)^{-1/2}$ singularity in the time-domain, rather than a $(t^2-r^2)^{-1/2}$, the relative differences only becoming large when $t >> r$.

3.6 Green's Function Solution to Maxwell's Equation

In Chapter 2, a gauge transform, together with the Lorentz condition, was used to solve Maxwell's equations and reduce them to two independent time dependent wave equations given by

$$\nabla^2 U - \frac{1}{c^2}\frac{\partial^2 U}{\partial t^2} = -4\pi\rho$$

and

$$\nabla^2 \mathbf{A} - \frac{1}{c^2}\frac{\partial^2 \mathbf{A}}{\partial t^2} = -\frac{4\pi}{c}\mathbf{j}.$$

Having discussed the time dependent Green's functions for the wave equation, we can now investigate the general solution to Maxwell's equations under the Lorentz condition. In particular, we consider the solution for the electric scalar potential U given ρ. The form of analysis is the same as used before, throughout this Chapter. Thus, solving for U, using Green's theorem (with homogeneous boundary conditions) and the conditions that u and $\partial u/\partial t$ are zero at $t = \pm\infty$ gives

$$U(\mathbf{r}_0, t_0) = \int_{-\infty}^{\infty}\int 4\pi\rho(\mathbf{r}, t)G(\mathbf{r} \mid \mathbf{r}_0, t \mid t_0)d^3\mathbf{r}dt = \int d^3\mathbf{r}\int_{-\infty}^{\infty} dt\frac{\rho(\mathbf{r}, t)}{R}\delta\left(\frac{R}{c} + t - t_0\right)$$

$$= \int d^3\mathbf{r}\frac{\rho\left(\mathbf{r}, t_0 - \frac{R}{c}\right)}{R}$$

where $R = \mid \mathbf{r} - \mathbf{r}_0 \mid$ or

$$U(\mathbf{r}_0, t_0) = \int \frac{\rho(\mathbf{r}, \tau)}{R}d^3\mathbf{r}$$

where

$$\tau = t_0 - \frac{R}{c}.$$

The solution for the Magnetic Vector Potential $\mathbf{A}$ can be found by solving for the components A_x, A_y and A_z separately. These are all scalar equations of exactly the same type and therefore have identical solutions and combine to give

$$\mathbf{A}(\mathbf{r}_0, t_0) = \int \frac{\mathbf{j}(\mathbf{r}, \tau)}{cR}d^3\mathbf{r}$$

The wavefields U and $\mathbf{A}$ are called the Retarded Potentials. The current value of U at $(\mathbf{r}_0, t_0)$ depends on ρ at earlier times $\tau = t_0 - R/c$. A change in ρ or $\mathbf{j}$ affects U and $\mathbf{A}$ (and hence $\mathbf{e}$ and $\mathbf{b}$) R/c seconds later - the change propagates outward at velocity c. This is the principle of electromagnetic wave propagation.

3.7 Green's Function for the Diffusion Equation

We calculate the Green's function G for the diffusion equation

$$\nabla^2 u(\mathbf{r},t) = \sigma \frac{\partial}{\partial t} u(\mathbf{r},t), \;\; \sigma = \frac{1}{D}$$

where D is the 'Diffusivity', satisfying the homogeneous boundary conditions and the causality condition

$$G(\mathbf{r} \mid \mathbf{r}_0, t \mid t_0) = 0 \;\text{ if }\; t < t_0.$$

This can be accomplished for one-, two- and three-dimensions simultaneously. Thus with $R = \mid \mathbf{r} - \mathbf{r}_0 \mid$ and $\tau = t - t_0$ we require the solution of the equation

$$\left(\nabla^2 - \sigma \frac{\partial}{\partial \tau}\right) G(R,\tau) = -\delta^n(R)\delta(\tau), \;\; \tau > 0$$

where n is 1, 2 or 3 depending on the number of dimensions. One way of solving this equation is first to take the Laplace transform with respect to τ, then solve for G (in Laplace space) and then inverse Laplace transform the result. This requires an initial condition to be specified (the value of G at $\tau = 0$). Another way to solve this equation is to take its Fourier transform with respect to R, solve for G (in Fourier space) and then inverse Fourier transform the result. Here, we adopt the latter approach. Let

$$G(R,\tau) = \frac{1}{(2\pi)^n} \int_{-\infty}^{\infty} \widetilde{G}(\mathbf{k},\tau) \exp(i\mathbf{k}\cdot\mathbf{R}) d^n\mathbf{k}$$

and

$$\delta^n(R) = \frac{1}{(2\pi)^n} \int_{-\infty}^{\infty} \exp(i\mathbf{k}\cdot\mathbf{R}) d^n\mathbf{k}.$$

Then the equation for G reduces to

$$\sigma \frac{\partial \widetilde{G}}{\partial \tau} + k^2 \widetilde{G} = \delta(\tau)$$

which has the solution

$$\widetilde{G} = \frac{1}{\sigma} \exp(-k^2\tau/\sigma) H(\tau)$$

where $H(\tau)$ is the step function

$$H(\tau) = \begin{cases} 1, & \tau > 0; \\ 0, & \tau < 0. \end{cases}$$

Hence, the Green's functions are given by

$$G(R,\tau) = \frac{1}{\sigma(2\pi)^n} H(\tau) \int_{-\infty}^{\infty} \exp(i\mathbf{k}\cdot\mathbf{R}) \exp(-k^2\tau/\sigma) d^n\mathbf{k}$$

$$= \frac{1}{\sigma(2\pi)^n} H(\tau) \left(\int_{-\infty}^{\infty} \exp(ik_x R_x) \exp(-k_x^2 \tau/\sigma) dk_x \right) ...$$

By rearranging the exponent in the integral, it becomes possible to evaluate each integral exactly. Thus, with

$$ik_x R_x - k_x^2 \frac{\tau}{\sigma} = -\left(k_x \sqrt{\frac{\tau}{\sigma}} - i \frac{R_x}{2} \sqrt{\frac{\sigma}{\tau}} \right)^2 - \left(\frac{\sigma R_x^2}{4\tau} \right) = -\frac{\tau}{\sigma} \xi^2 - \left(\frac{\sigma R_x^2}{4\tau} \right)$$

where

$$\xi = k_x - i \frac{\sigma R_x}{2\tau}.$$

the integral over k_x becomes

$$\int_{-\infty}^{\infty} \exp\left[-\left(\frac{\tau}{\sigma} \xi^2 \right) - \left(\frac{\sigma R_x}{4\tau} \right) \right] d\xi = e^{-(\sigma R_x^2/4\tau)} \int_{-\infty}^{\infty} e^{-(\tau \xi^2/\sigma)} d\xi$$

$$= \sqrt{\frac{\pi\sigma}{\tau}} \exp\left[-\left(\frac{\sigma R_x^2}{4\tau} \right) \right]$$

with similar results for the integrals over k_y and k_z giving the result

$$G(R, \tau) = \frac{1}{\sigma} \left(\frac{\sigma}{4\pi\tau} \right)^{\frac{n}{2}} \exp\left[-\left(\frac{\sigma R^2}{4\tau} \right) \right] H(\tau).$$

The function G satisfies an important property which is valid for all n:

$$\int_{-\infty}^{\infty} g(R, \tau) d^n \mathbf{r} = \frac{1}{\sigma}; \quad \tau > 0.$$

This is the expression for the conservation of the Green's function associated with the diffusion equation. If, at a time t_0 and at a point in space $\mathbf{r}_0$, a source is introduced which starts to diffuse, then the diffusion process through the medium characterized by σ is such that the total flux is unchanged.

3.8 Green's Functions for the Laplace and Poisson Equations

The Laplace and Poisson equations (in one- or two-dimensions) are given by

$$\nabla^2 u = 0$$

and

$$\nabla^2 u = -f$$

respectively. Let us consider the Poisson equation first. The general approach is identical to that used to derive a solution to the inhomogeneous wave equation.

Thus, working in three dimensions and defining the Green's function to be the solution of

$$\nabla^2 g(\mathbf{r} \mid \mathbf{r}_0) = -\delta^3(\mathbf{r} - \mathbf{r}_0)$$

from Poisson's equation, we obtain the following result

$$u = \oint_S (g\nabla u - u\nabla g) \cdot \hat{\mathbf{n}} d^2\mathbf{r} + \int_V g f d^3\mathbf{r}$$

where we have used Green's theorem to obtain the surface integral on the right hand side. The problem now is to find the Green's function for this problem. Clearly, since the solution to the equation

$$(\nabla^2 + k^2)g = -\delta^3(\mathbf{r} - \mathbf{r}_0)$$

is

$$g(\mathbf{r} \mid \mathbf{r}_0, k) = \frac{1}{4\pi \mid \mathbf{r} - \mathbf{r}_0 \mid} \exp(ik \mid \mathbf{r} - \mathbf{r}_0 \mid)$$

we should expect the Green's function for the three-dimensional Poisson equation (and the Laplace equation) to be of the form

$$g(\mathbf{r} \mid \mathbf{r}_0) = \frac{1}{4\pi \mid \mathbf{r} - \mathbf{r}_0 \mid}.$$

This can be shown by taking the Fourier transform of the equation for g which gives

$$k^2 G(k) = 1$$

where

$$G(k) = \int g(R) \exp(i\mathbf{k} \cdot \mathbf{R}) d^3\mathbf{R}, \quad R = \mid \mathbf{r} - \mathbf{r}_0 \mid .$$

Therefore

$$g(R) = \frac{1}{(2\pi)^3} \int \frac{\exp(i\mathbf{k} \cdot \mathbf{R})}{k^2} d^3\mathbf{k} = \frac{1}{(2\pi)^3} \int_0^{2\pi} d\phi \int_{-1}^{1} d(\cos\theta) \int_0^{\infty} dk \exp(ikR\cos\theta)$$

$$= \frac{1}{2\pi^2 R} \int_0^{\infty} \frac{\sin(kR)}{k} dk = \frac{1}{4\pi R}$$

using spherical polar coordinates and the result

$$\int_0^{\infty} \frac{\sin x}{x}\, dx = \frac{\pi}{2}.$$

Thus, we obtain the following fundamental result:

$$\nabla^2 \left(\frac{1}{4\pi R} \right) = -\delta^3(R).$$

TABLE 2 Outgoing Free space Green's functions for the wave equation in one-, two- and three-dimensions.

Dimension	Time-ependent Function	Time-independent Function
1D	$\frac{c}{4}\text{sgn}(\tau - R/c)$	$\frac{i}{2k}\exp(ikR) = \frac{i}{2k},\ \ kR \to 0$
2D	$\frac{1}{4\pi\sqrt{\tau^2 - R^2}}$	$\frac{i}{4}H_0^{(1)}(kR)$ $= \frac{\exp(ikR)}{\sqrt{kR}},\ \ kR \to \infty$ (ignoring scaling) $= -\ln(kR),\ \ kR \to 0$ (ignoring scaling)
3D	$\frac{\delta(\tau + R/c)}{4\pi R}$	$\frac{\exp(ikR)}{4\pi R} = \frac{1}{4\pi R},\ \ kR \to 0$

With homogeneous boundary conditions, the solution to the Poisson equation is

$$u(\mathbf{r}_0) = \frac{1}{4\pi}\int_V \frac{f(\mathbf{r})}{\mid \mathbf{r} - \mathbf{r}_0 \mid} d^3\mathbf{r}.$$

In two dimensions the solution is of the same form, but with a Green's function given by

$$g(\mathbf{r} \mid \mathbf{r}_0) = \frac{1}{2\pi}\ln \mid \mathbf{r} - \mathbf{r}_0 \mid .$$

Clearly, the general solution to Laplace's equation (in 3D) is

$$u = \oint_S (g\nabla u - u\nabla g) \cdot \hat{\mathbf{n}} d^2\mathbf{r}.$$

These solutions to the Laplace and Poisson equations are analogous to those for the homogeneous and inhomogeneous wave equations. The principle behind the method of solution is the same; what changes is the Green's function.

3.9 Discussion

This chapter has been designed to provide an introduction to the use of Green's functions for solving partial differential equations in different dimensions and for time dependent and time independent problems. The material presented has been based almost exclusively on the use of free space Green's functions in which a solution is developed over the infinite domain to which boundary conditions can be applied. The focus has been on Green's functions for wave equations as this is the principal basis for modelling imaging systems and image understanding. The free-space Green's functions for the wave equation are summarised in Table 2 (where $R \equiv \mid \mathbf{r} - \mathbf{r}_0 \mid$)

4 ELECTROMAGNETIC SCATTERING THEORY

A scattered wavefield depends on both the nature of the scatterer and the type and properties of the radiation scattered by it. These properties are described by the characteristic inhomogeneous wave equations. Chapter 2 introduced the field equations and wave equations that form a basis for modelling electromagnetic scattering and and, in turn, EM imaging systems. For scalar wavefields u, it was shown that we can derive inhomogeneous wave equations of the form

$$(\nabla^2 + k^2)u = -\hat{L}u$$

or for vector fields $\mathbf{u}$ of the form

$$(\nabla^2 + k^2)\mathbf{u} = -\hat{L}\mathbf{u}.$$

For a non-conductive linear isotropic electromagnetic scatterer with variations in the permittivity γ_ϵ and permeability γ_μ,

$$(\nabla^2 + k^2)\widetilde{\mathbf{E}} = -\hat{L}\widetilde{\mathbf{E}}$$

with

$$\hat{L}\widetilde{\mathbf{E}} = k^2\gamma_\epsilon\widetilde{\mathbf{E}} + \nabla(\widetilde{\mathbf{E}} \cdot \nabla \ln\epsilon) + \nabla \times (\gamma_\mu \nabla \times \widetilde{\mathbf{E}}).$$

In this Chapter, we explore the use of the Green's function for solving inhomogeneous wave equations of the type

$$(\nabla^2 + k^2)u = -k^2\gamma u$$

which is known as the inhomogeneous Helmholtz equation. This problem is related to volume scattering when $\gamma(\mathbf{r})$ is of compact support (i.e. $\mathbf{r} \in V$) We also study the solutions to the homogeneous Helmholtz equation

$$(\nabla^2 + k^2)u = 0$$

where u and ∇u are defined on a boundary defining a surface which generates surface scattering. The homogeneous and inhomogeneous Helmholtz equations provide the

basis for developing a scattering theory that is of value in imaging science. We can then investigate scattering models that are based on physical models (inhomogeneous wave equations) that are more complete (i.e. models that describe a greater number of physical effects) when the right hand side of the wave equations considered here is of the form $-\hat{L}u$. The methods used are explicitly based on application of Green's functions presented in Chapter3.

Much of the original work on scattering theory began in the 1930s and has been the product of mathematicians and physicists working on problems of theoretical physics including quantum mechanics and high energy nuclear physics, where the scattering of particles and the interpretation of their 'images' (particle tracking devices) has been fundamental to investigating the structure of matter.

4.1 The Inhomogeneous Helmholtz Equation

The inhomogeneous Helmholtz equation is given by

$$(\nabla^2 + k^2)u(\mathbf{r}, k) = -k^2\gamma(\mathbf{r})u(\mathbf{r}, k)$$

where γ is an inhomogeneity which is responsible for scattering the wavefield u and is therefore sometimes referred to as a scatterer - usually considered to be of compact support. In electromagnetism, the Helmholtz equation is obtained when we can assume that the medium is non conductive (i.e. $\sigma = 0$), γ_μ is a constant and the term $\nabla(\mathbf{E} \cdot \nabla \ln\epsilon)$ is ignored.

The Helmholtz equation can be derived quite generally from the time dependent wave equation

$$\left(\nabla^2 - \frac{1}{c^2}\frac{\partial^2}{\partial t^2}\right) U(\mathbf{r}, t) = 0$$

by letting [14]

$$\frac{1}{c^2} = \frac{1}{c_0^2}(1 + \gamma)$$

where γ is a dimensionless quantity and c_0 is a constant (wave speed). Note that the form of the wave equation dictates that c must be of finite value. If a wavefield (whatever the field may be) was to convey information from one point in space to another instantaneously then the second term would be zero and the 'wave equation' would be reduced to 'Laplace's equation' and the independent variable t would become an irrelevance! The upper limit at which any wavefield can propagate is determined by the speed of an electromagnetic wave in a perfect vacuum. However, in a more general perspective, the rationale associated with the fact that c **must** be finite (as given above) means that the influence of any physical field (whether it be an electric, magnetic, gravitational, weak or strong force field) on any measurable entity can only occur in a finite period of time and that there can be no such thing as instantaneous 'action at a distance'. This is the essential difference between the 'universe' according to Isaac Newton and the 'universe' according to Albert Einstein,

a difference that, given the wave equation, points to the 'physics' of a wavefield being more fundamental than the 'physics' of the field itself. This principle should be considered in light of the fact that the one property common to all the principal field equation of physics (i.e. Einstein's equations, Maxwell's equations and Dirac's equations), is that they describe wave phenomena (i.e. gravity wave, electromagnetic wave and matter waves respectively).

With $U(\mathbf{r},t) = u(\mathbf{r},\omega)\exp(i\omega t)$ we have

$$(\nabla^2 + k^2)u(\mathbf{r},k) = -k^2\gamma(\mathbf{r})u(\mathbf{r},k)$$

where

$$k = \frac{\omega}{c_0}.$$

Note that we can write the Schrödinger equation in terms of the Helmholtz equation since, from the postulates of quantum mechanics,

$$\frac{1}{c^2} = \frac{1}{c_0^2}(1+\gamma) = \frac{2m(E-E_p)}{E^2}$$

where E is the energy of a particle of mass m subject to a potential with potential energy E_p and thus, the Schrödinger equation is

$$(\nabla^2 + k^2)u(\mathbf{r},k) = -k^2\gamma(\mathbf{r})u(\mathbf{r},k)$$

with

$$\gamma(\mathbf{r}) = 2mc_0^2\frac{[E-E_p(\mathbf{r})]}{E^2} - 1.$$

Note that since γ is dimensionless, this result implies that mc_0^2 is energy (i.e. Einstein's famous energy-mass equivalence formula, where c_0 is the speed of light and m is the rest mass). Thus, for a scalar electromagnetic wavefield interacting with a non-conductive dielectric - ignoring the term $\nabla(\mathbf{E}\cdot\nabla\ln\epsilon_r)$ the Helmholtz equation is given by

$$(\nabla^2 + k^2)u(\mathbf{r},k) = -k^2\gamma(\mathbf{r})u(\mathbf{r},k)$$

where $\gamma = \epsilon_r - 1$. We can therefore interpret the relative permittivity ϵ_r in terms of the function

$$2mc_0^2\frac{[E-E_p(\mathbf{r})]}{E^2}$$

on an entirely phenomenological basis. We now consider solutions to the inhomogeneous Helmholtz equation

4.2 Solutions to the Helmholtz Equation

The Green's function method presented in Chapter 3 can be used to solve the inhomogeneous Helmholtz equation. The basic solution is (under the assumption that γ is of compact support $\mathbf{r} \in V$) [15], [16], [17]

$$u(\mathbf{r}_0,k) = k^2\int_V g\gamma u d^3\mathbf{r} + \oint_S (g\nabla u - u\nabla g)\cdot\hat{\mathbf{n}}d^2\mathbf{r}.$$

To compute the surface integral, a condition for the behaviour of u on the surface S of γ must be chosen. Consider the case where the incident wavefield u_i is a simple plane wave of unit amplitude

$$\exp(i\mathbf{k}\cdot\mathbf{r})$$

satisfying the homogeneous wave equation

$$(\nabla^2 + k^2)u_i(\mathbf{r},k) = 0.$$

By choosing the condition $u(\mathbf{r},k) = u_i(\mathbf{r},k)$ on the surface of γ, we obtain the result

$$u(\mathbf{r}_0,k) = k^2 \int_V g\gamma u d^3\mathbf{r} + \oint_S (g\nabla u_i - u_i\nabla g)\cdot\hat{\mathbf{n}}d^2\mathbf{r}.$$

Now, using Green's theorem to convert the surface integral back into a volume integral, we have

$$\oint_S (g\nabla u_i - u_i\nabla g)\cdot\hat{\mathbf{n}}d^2\mathbf{r} = \int_V (g\nabla^2 u_i - u_i\nabla^2 g)d^3\mathbf{r}.$$

Noting that

$$\nabla^2 u_i = -k^2 u_i$$

and that

$$\nabla^2 g = -\delta^3 - k^2 g$$

we obtain

$$\int_V (g\nabla^2 u_i - u_i\nabla^2 g)d^3\mathbf{r} = \int \delta^3 u_i d^3\mathbf{r} = u_i.$$

Hence, by choosing the field u to be equal to the incident wavefield u_i on the surface of γ, we obtain a solution of the form

$$u = u_i + u_s$$

where

$$u_s = k^2 \int_V g\gamma u d^3\mathbf{r}.$$

The function u_s is the scattered wavefield and the above equation is generally known as the Lippmann-Schwinger equation [14].

4.2.1 The Born Approximation

From the last result it is clear that, in order to compute the scattered field u_s, we must define u inside the volume integral. Unlike the surface integral, a boundary condition will not help here because it is not sufficient to specify the behaviour of u at a boundary. In this case, the behaviour of u throughout V needs to be known. In general, it is not possible to do this (i.e. to compute the scattered wavefield exactly) and we are forced to choose a model for u inside V that is compatible with a particular physical problem in the same way that an appropriate set of boundary

conditions is required to evaluate the surface integral. The simplest model for the internal field is based on assuming that u behaves like u_i for $\mathbf{r} \in V$. The scattered field is then given by

$$u_s(\mathbf{r}_0, k) = k^2 \int_V g(\mathbf{r} \mid \mathbf{r}_0, k)\gamma(\mathbf{r})u_i(\mathbf{r}, k)d^3\mathbf{r}.$$

This assumption provides an approximate solution for the scattered field. It is known as the Born approximation after Max Born who first introduced it in his study of quantum mechanics in the 1930s.

There is another way of deriving this result that is instructive; it helps us to obtain a criterion for the validity of this approximation which will be considered shortly. We start with the inhomogeneous Helmholtz equation

$$(\nabla^2 + k^2)u = -k^2\gamma u$$

and consider a solution for u in terms of a sum of the incident and scattered fields, i.e.

$$u = u_i + u_s.$$

The wave equation then becomes

$$(\nabla^2 + k^2)u_s + (\nabla^2 + k^2)u_i = -k^2\gamma(u_i + u_s).$$

If the incident field satisfies

$$(\nabla^2 + k^2)u_i = 0,$$

then

$$(\nabla^2 + k^2)u_s = -k^2\gamma(u_i + u_s).$$

Assuming that

$$u_i + u_s \simeq u_i, \quad \mathbf{r} \in V$$

we obtain

$$(\nabla^2 + k^2)u_s \simeq -k^2\gamma u_i, \quad \mathbf{r} \in V.$$

Solving for u_s and using the homogeneous boundary conditions (i.e. $u_s = 0$ on S and $\nabla u_s = 0$ on S) we obtain

$$u_s = \oint_S (g\nabla u_s - u_s\nabla g) \cdot \hat{\mathbf{n}} d^2\mathbf{r} + k^2 \int_V g\gamma u_i d^3\mathbf{r} = k^2 \int_V g\gamma u_i d^3\mathbf{r}.$$

4.2.2 Validity of the Born Approximation

In general, the Born approximation requires that u_s is 'small' compared to u_i. What do we really mean by the term 'small' and how can we quantify it? One way to answer this question is to compute an appropriate measure for both the incident and scattered fields and compare the two results. Consider the case where we compute the root mean square modulus (i.e. the L_2 norm) of each field. We then require that

$$\left(\int_V \mid u_s(\mathbf{r}_0, k) \mid^2 d^3\mathbf{r}_0 \right)^{\frac{1}{2}} << \left(\int_V \mid u_i(\mathbf{r}_0, k) \mid^2 d^3\mathbf{r}_0 \right)^{\frac{1}{2}}$$

or[1]

$$\frac{\|u_s\|}{\|u_i\|} << 1. \tag{4.1}$$

Essentially, this condition means that the overall intensity of u_s in V is small compared to that of u_i in V.

Let us now look in more detail at the nature of this condition. Ideally, what we want is a version of the condition that can be cast in terms of a set of physical parameters (such as the wavelength and the physical extent of γ for example). The Born scattered field at $\mathbf{r}_0$ is given by

$$u_s(\mathbf{r}_0, k) = k^2 \int_V g(\mathbf{r} \mid \mathbf{r}_0, k)\gamma(\mathbf{r})u_i(\mathbf{r}, k)d^3\mathbf{r}.$$

By taking the L_2 norm of this equation we can write

$$\|u_s(\mathbf{r}_0, k)\| = \|k^2 \int_V g(\mathbf{r} \mid \mathbf{r}_0, k)\gamma(\mathbf{r})u_i(\mathbf{r}, k)d^3\mathbf{r}\|$$

$$\leq k^2\|u_i(\mathbf{r}_0, k)\| \times \| \int_V g(\mathbf{r} \mid \mathbf{r}_0, k)\gamma(\mathbf{r})d^3\mathbf{r}\|.$$

Using this result, the condition required for the Born approximation to hold [i.e. condition (4.1)] can be written as

$$k^2\| \int_V g(\mathbf{r} \mid \mathbf{r}_0, k)\gamma(\mathbf{r})d^3\mathbf{r}\| << 1, \quad \mathbf{r}_0 \in V. \tag{4.2}$$

Here, the norm involves integration over the spatial variable $\mathbf{r}_0$ in the scattering volume V. To emphasize this we write $\mathbf{r}_0 \in V$.

Condition (4.2) can be written as

$$I(\mathbf{r}_0) << 1$$

where

$$I(\mathbf{r}_0) = k^2\| \int_V g(\mathbf{r} \mid \mathbf{r}_0, k)\gamma(\mathbf{r})d^3\mathbf{r}\|$$

$$\leq k^2 \left(\int_V \mid g(\mathbf{r} \mid \mathbf{r}_0, k) \mid^2 d^3\mathbf{r} \right)^{\frac{1}{2}} \left(\int_V \mid \gamma(\mathbf{r}) \mid^2 d^3\mathbf{r} \right)^{\frac{1}{2}}.$$

Substituting the expression for the three-dimensional Green function into the above expression, we have

$$I(\mathbf{r}_0) \leq k^2 \left(\frac{1}{16\pi^2} \int_V \frac{1}{\mid \mathbf{r} - \mathbf{r}_0 \mid^2} d^3\mathbf{r} \int_V \mid \gamma(\mathbf{r}) \mid^2 d^3\mathbf{r} \right)^{\frac{1}{2}}.$$

[1] where $\| \bullet \|$ is taken to denote the L_2 norm, i.e. $\| \bullet \|_2$

A relatively simple calculation can now be performed, if we consider γ to be a sphere of volume V and radius R, and resort to calculating its least upper bound which occurs when $\mathbf{r}_0 = 0$. Using spherical polar coordinates (r, θ, ϕ), we have

$$\sup \int_V \frac{1}{\mid \mathbf{r} - \mathbf{r}_0 \mid^2} d^3\mathbf{r} = \int_V \frac{1}{r^2} d^3\mathbf{r} = \int_0^{2\pi} \int_{-1}^{1} \int_0^R dr d(\cos\theta) d\phi = 4\pi R$$

where **sup** denotes the supremum. Using this result, we can write

$$\sup I(\mathbf{r}_0) \leq k^2 \left(\frac{R}{4\pi} \int_V \mid \gamma(\mathbf{r}) \mid^2 d^3\mathbf{r} \right)^{\frac{1}{2}}$$

and noting that

$$V = \int_V d^3\mathbf{r} = \frac{4}{3}\pi R^3$$

we obtain

$$\sup I(\mathbf{r}_0) \leq \frac{1}{\sqrt{3}} k^2 R^2 \bar{\gamma}$$

where

$$\bar{\gamma} = \sqrt{\frac{\int \mid \gamma \mid^2 d^3\mathbf{r}}{\int d^3\mathbf{r}}}.$$

Hence, the condition for the Born approximation to apply becomes (ignoring $\sqrt{3}$)

$$k^2 R^2 \bar{\gamma} << 1$$

or

$$\bar{\gamma} << \frac{1}{k^2 R^2}.$$

This condition demonstrates that, in principle, large values of γ can occur so long as its root mean square value over the volume V is small compared to $1/k^2R^2$. In scattering theory, γ is said to be a 'weak scatterer'. Note that when k or R approaches zero, this condition is easy to satisfy. Born scattering is thus, more likely to occur in situations when

$$\frac{\lambda}{R} >> 1$$

where λ is the wavelength (noting that $k = 2\pi/\lambda$). If

$$\frac{\lambda}{R} \sim 1$$

then the value of $\bar{\gamma}$ must be small for Born scattering to occur.

By repeating the method given above and using the two- and one-dimensional Green functions, respectively, it is easy to show that in two dimensions the condition required for the Born approximation to apply is given by

$$\bar{\gamma} << \frac{1}{(kR)^{3/2}}$$

where R is the radius of a disc of area A and $\bar{\gamma}$ is the root mean square over A.

In one dimension, the result is

$$\bar{\gamma} << \frac{1}{kL}$$

where L is the length of the scatterer and $\bar{\gamma}$ is the root mean square over L. In both cases we use the same Green function solution to solve the 2D and 1D inhomogeneous Helmholtz equations, respectively. In each case, we assume that the scattered field is, on average, weak compared to the incident field. We may consider the term 'weak' to imply that the total energy associated with u_s inside the inhomogeneity γ is small compared to u_i outside the scatterer.

4.2.3 Asymptotic Born Scattering

By measuring u_s, we can attempt to invert the relevant integral equation and hence recover or reconstruct γ. This type of problem is known as the inverse scattering problem, and solutions to this problem are called inverse scattering solutions. This subject is one of the most fundamental in mathematical physics and is the subject of continuing research. The simplest type of inverse scattering problem occurs when a Born scattered wavefield is measured in the far field or Fraunhofer zone (i.e. when the Green functions takes on its asymptotic form discussed in Chapter 3). From previous results, working in 3D, when the incident field is a (unit) plane wave

$$u_i = \exp(ik\hat{\mathbf{n}}_i \cdot \mathbf{r})$$

where $\hat{\mathbf{n}}_i$ points in the direction of the incident field, the Born scattered field observed at $\mathbf{r}_s$ is

$$u_s(\hat{\mathbf{n}}_s, \hat{\mathbf{n}}_i, k) = \frac{k^2}{4\pi r_s} \exp(ikr_s) \int_V \exp[-ik(\hat{\mathbf{n}}_s - \hat{\mathbf{n}}_i) \cdot \mathbf{r}] \gamma(\mathbf{r}) d^3\mathbf{r}, \quad \mathbf{r} \in V$$

where $\hat{\mathbf{n}}_s (= \mathbf{r}_s / r_s)$ denotes the direction in which u_s propagates. From this result, it is clear that the function γ can be recovered from u_s by three-dimensional Fourier inversion. The scattered field produced by a two-dimensional Born scatterer in the far field is given by

$$u_s(\hat{\mathbf{n}}_i, \hat{\mathbf{n}}_s, k) = \frac{\exp(i\pi/4)}{\sqrt{8\pi}} \frac{k^2}{\sqrt{kr_s}} \exp(ikr_s) \int_A \exp[-ik(\hat{\mathbf{n}}_s - \hat{\mathbf{n}}_i) \cdot \mathbf{r}] \gamma(\mathbf{r}) d^2\mathbf{r}, \quad \mathbf{r} \in A.$$

In one dimension, the equivalent result is (for a right travelling wave)

$$u_s(x_s, k) = \frac{ik}{2} \exp(ikx_s) \int_L \gamma(x) dx, \quad x \in L.$$

When $\hat{\mathbf{n}}_s = \hat{\mathbf{n}}_i$, we see that

$$u_s = \frac{k^2}{4\pi r_s} \exp(ikr_s) \int_V \gamma(\mathbf{r}) d^3\mathbf{r}.$$

This is called the forward-scattered field. In terms of Fourier analysis, it represents the zero frequency or DC component of the spectrum of γ. Another special case arises when $\hat{\mathbf{n}}_s = -\hat{\mathbf{n}}_i$. The scattered field that is produced in this case is called the back-scattered field, and in three dimensions is given by

$$u_s(\hat{\mathbf{n}}_s, k) = \frac{k^2}{4\pi r_s} \exp(ikr_s) \int_V \exp(-2ik\hat{\mathbf{n}}_s \cdot \mathbf{r})\gamma(\mathbf{r})d^3\mathbf{r}.$$

In one dimension, the result is (for a left travelling wave)

$$u_s(k) = \frac{ik}{2} \exp(ikx_s) \int_L \exp(-2ikx)\gamma(x)dx.$$

Note that, in one-dimension, the scattering function can only be recovered (via Fourier inversion) by measuring the back-scattered spectrum whereas in two and three dimensions, the scattering function can, in principle, be recovered by either keeping k fixed or varying k.

4.3 Examples of Born Scattering

By way of a short introduction to the applications and uses of the Born approximation, some well known examples are now presented in which it is used to derive expressions for the scattered intensity associated with three physically different scattering phenomena - Rutherford scattering, Rayleigh scattering and Tyndall.

4.3.1 Rutherford Scattering

Rutherford scattering ranks as one of the most important experiments of the Twentieth Century because it was the basis for developing the basic 'visual model' for the atom - a positively charged nucleus with negatively charged orbiting electrons.

In Rutherford's famous experiment (which dates from 1910), α-particles (or helium nuclei) were scattered by gold leaf. The differential cross-section denoted by $d\sigma/d\Omega$ (i.e. the number of particles scattered into a solid angle $d\Omega$ per unit time divided by the number of particles incident per unit area per unit time) was then measured at different scattering angles θ. By treating the α-particles as classical Newtonian particles, Rutherford showed that if the scattering potential (i.e. due to the nucleus of the atoms in the gold leaf) is a repulsive Coulomb potential, then

$$\frac{d\sigma}{d\Omega} \propto \frac{1}{\sin^4(\theta/2)}.$$

This was before the development of quantum mechanics and the emergence of Schrödinger's equation as a governing partial differential equation of quantum mechanics. In this Section, we shall derive Rutherford's result by solving Schrödinger's equation using a Green function.

In terms of quantum mechanics we can consider Rutherford's scattering experiment to consist of a source of plane waves (i.e. the de Broglie or probability waves associated with the α-particles), a scattering function denoted by E_p (the potential associated with the nucleus of the atoms which make up the gold leaf) and a measuring device which allows us to record the intensity of the scattered radiation at different angles to the incident beam. The Green function solution to the 3D Schrödinger equation

$$(\nabla^2 + k^2)u(\mathbf{r}, k) = V(\mathbf{r})u(\mathbf{r})$$

for an incident plane wave $u_i(\mathbf{r}, k) = \exp(i\mathbf{k} \cdot \mathbf{r})$ is given by

$$u(\mathbf{r}_0, k) = u_i(\mathbf{r}_0, k) + \int g(\mathbf{r} \mid \mathbf{r}_0, k)V(\mathbf{r})u(\mathbf{r}, k)d^3\mathbf{r}.$$

This is the Lippmann-Schwinger equation. The limits of the integral are left 'open' because this equation applies to potentials that are finite (of compact support) or asymptotic (tend to zero at infinity). The inversion of this integral equation is the basis for inverse Schrödinger scattering in three-dimensions.

The Born scattered wave in the far field due to a scattering potential E_p which is influential over all space is given by

$$u_s(\hat{\mathbf{n}}_s, \hat{\mathbf{n}}_i, k) = -\frac{\exp(ikr_s)}{4\pi r_s} \int_{-\infty}^{\infty} \exp[-ik(\hat{\mathbf{n}}_s - \hat{\mathbf{n}}_i) \cdot \mathbf{r}]V(\mathbf{r})d^3\mathbf{r}.$$

For fixed k and r_s (the distance at which the scattered wavefield is measured from the scattering event), the measured intensity I of the scattered wavefield is given by

$$I = u_s u_s^* = \frac{1}{16\pi^2 r_s^2} \mid A \mid^2$$

where A is the scattering amplitude,

$$A(\hat{\mathbf{n}}_s, \hat{\mathbf{n}}_i, k) = \int_{-\infty}^{\infty} \exp[-ik(\hat{\mathbf{n}}_s - \hat{\mathbf{n}}_i) \cdot \mathbf{r}]V(\mathbf{r})d^3\mathbf{r}.$$

The differential cross section measures the flux of particles through a given area in specific period of time. It is thus a measure of the wavefield intensity, i.e.

$$\frac{d\sigma}{d\Omega} = I.$$

Hence, using quantum mechanics (i.e. Schrödinger's equation), the differential cross-section for Rutherford's scattering experiment can be obtained by evaluating the Fourier transform of the potential E_p. For a radially symmetric potential $E_p(r)$, the scattering amplitude becomes (switching to spherical polar coordinates r, ϕ, ψ)

$$A(\hat{\mathbf{n}}_s, \hat{\mathbf{n}}_i) = \int_0^{2\pi} d\psi \int_{-1}^{1} d(\cos\phi) \int_0^{\infty} dr \;\; r^2 \exp(-ik \mid \hat{\mathbf{n}}_s - \hat{\mathbf{n}}_i \mid r\cos\phi)V(r).$$

The modulus of $\hat{\mathbf{n}}_s - \hat{\mathbf{n}}_i$ is given by

$$\mid \hat{\mathbf{n}}_s - \hat{\mathbf{n}}_i \mid = \sqrt{(\hat{\mathbf{n}}_s - \hat{\mathbf{n}}_i) \cdot (\hat{\mathbf{n}}_s - \hat{\mathbf{n}}_i)} = \sqrt{2(1 - \cos\theta)}$$

where

$$\cos\theta = \hat{\mathbf{n}}_s \cdot \hat{\mathbf{n}}_i$$

and θ is the scattering angle (the angle between the incident and scattered fields). Using the half angle formula,

$$1 - \cos\theta = 2\sin^2(\theta/2)$$

we can write

$$\mid \hat{\mathbf{n}}_s - \hat{\mathbf{n}}_i \mid = 2\sin(\theta/2)$$

and integrating over ψ and $\cos\phi$ the scattering amplitude as a function θ can be written as

$$A(\theta) = \frac{2\pi}{k\sin(\theta/2)} \int_0^\infty \sin[2kr\sin(\theta/2)] V(r) r dr.$$

All we need to do now is compute the remaining integral over r. If we use a simple Coulomb potential where $E_p(r) \propto 1/r$, then we run into a problem because the integrand does not converge as $r \to \infty$. For this reason, another radially symmetric potential is introduced which is given by

$$E_p(r) = \frac{\exp(-ar)}{r}$$

where $a > 0$ is a constant. This type of potential is known as a screened Coulomb potential, the parameter a determining the range over which the potential is influential. It allows us to evaluate the scattering amplitude analytically. We can then observe the behaviour of $\mid A \mid^2$ for a Coulomb potential by letting a approach zero. The scattering amplitude becomes

$$A(\theta) = \frac{2\pi}{k\sin(\theta/2)} \int_0^\infty \sin[2kr\sin(\theta/2)] \exp(-ar) dr.$$

This integral is given by

$$\frac{2k\sin(\theta/2)}{a^2 + [2k\sin(\theta/2)]^2}$$

and we can write

$$A(\theta) = \frac{\pi}{k^2\sin^2(\theta/2)} \left(1 + \frac{a^2}{[2k\sin(\theta/2)]^2}\right)^{-1}.$$

Hence, as $a \to 0$, we obtain

$$A(\theta) \simeq \frac{\pi}{k^2\sin^2(\theta/2)}$$

and the intensity of the scattered field is

$$I = \mid A(\theta) \mid^2 \propto \frac{1}{\sin^4(\theta/2)}.$$

We may think of Rutherford's scattering experiment as an inverse scattering problem in the sense that he deduced the potential of the nucleus by recording the way in which it scattered α-particles. However, he did not actually solve the inverse problem directly because he assumed that the scattering potential acted like a repulsive Coulomb potential *a priori* and justified this hypothesis later by showing that the theoretical and experimental results were compatible. One final and interesting point to note is that in order to undertake the experiment Rutherford required a very thin foil which was only a few atoms thick. Gold leaf was the best possible technical solution to this problem at the time. The reason for this was that the α-particles needed (on average) to scatter only from one nucleus in order to investigate the repulsive Coulomb potential theory. If a thicker foil had been used, the α-particles may have scattered from a number of atoms as they passed through it. Multiple scattering would have led to an indeterminacy in the results. It is important to note that the Born approximation used here to verify Rutherford's results using a Green function solution to Schrödinger's equation is consistent with the concept of single, or weak, scattering.

4.3.2 Rayleigh Scattering

Rayleigh scattering is the scattering of electromagnetic radiation by small dielectric scatterers. It is named after the English scientist Lord Rayleigh who was one of the Nineteenth Century's most prolific scientists and made contributions in many areas in mathematics, physics and chemistry, including some of the earliest studies on the scattering of light following the development of James Clerk Maxwell's theory of electromagnetism.

If we consider a scalar electromagnetic wave theory, then we can consider a wave equation of the form

$$(\nabla^2 + k^2)u(\mathbf{r}, k) = -k^2\gamma(\mathbf{r})u(\mathbf{r}, k), \quad \gamma = \epsilon_r - 1; \quad \mathbf{r} \in V$$

to describe the behaviour of the electric field u, where ϵ_r is the relative permittivity of a dielectric of compact support E_p. This is of course a highly idealized case, but it helps to provide another demonstration of Born scattering in a form that is pertinent to the use of Green functions for solving physically significant problems.

In the context of electromagnetic scattering problems, the Born approximation is sometimes referred to as the Rayleigh-Gan approximation - just a different name for an identical mathematical technique. Using this approximation, the asymptotic form of the the scattered electric field is given by

$$u_s(\hat{\mathbf{n}}_s, \hat{\mathbf{n}}_i, k) = \frac{k^2}{4\pi r_s} \exp(ikr_s) \int_V \exp[-ik(\hat{\mathbf{n}}_s - \hat{\mathbf{n}}_i) \cdot \mathbf{r}]\gamma(\mathbf{r}) d^3\mathbf{r}.$$

There are two important differences between this equation and its counterpart in quantum mechanics (i.e. the Schrödinger equation). First, the coefficient in front of the integral possesses a factor k^2. Second, the integral itself is over a finite volume of space V which is determined by the spatial extent of γ. In quantum mechanics, the influence of a potential may be 'felt' over all space so that the integral is over $\pm\infty$. This is an important distinction between scattering problems in quantum mechanics which involve asymptotic potentials (potentials which go to zero at infinity) and classical scattering problems of the type considered here.

Let us consider a model where a plane electromagnetic wave is incident on a homogeneous spherical dielectric object of radius R and relative permittivity ϵ_r. The theory which describes this type of scattering (scattering of light from uniform spheres) is called Mie theory. In this case the Born scattered amplitude is given by (following the same methods as those used earlier)

$$A(\theta) = \frac{2\pi k\gamma}{\sin(\theta/2)} \int_0^R \sin[2kr\sin(\theta/2)]rdr. \tag{4.3}$$

If the dimensions of the scatterer are small compared to the wavelength, then

$$kR << 1$$

and

$$\sin[2kr\sin(\theta/2)] \simeq 2kr\sin(\theta/2), \;\; 0 \leq r \leq R.$$

The scattering amplitude is then given by

$$A(\theta) \simeq 4\pi k^2\gamma \int_0^R r^2 dr = k^2\gamma V$$

where $V = 4\pi R^3/3$ is the volume of the scatterer. In this case, the scattering is entirely isotropic (i.e. the scattering amplitude is independent of the scattering angle). The intensity is proportional to k^4 or

$$\mid A(\theta) \mid^2 \;\propto\; \frac{1}{\lambda^4}.$$

Note the large inverse dependence on the wavelength. This result is characteristic of Rayleigh scattering and of the spectra produced by light scattering from small sub-wavelength structures. In the visible part of the spectrum, the intensity is greatest for blue light (the colour associated with the smallest wavelength of the visible spectrum). This is why the sky is blue, i.e. sunlight is scattered by the electrons in air molecules of the terrestrial atmosphere generating blue light preferentially around in all directions. Further, as the Sun approaches the horizon, we have to look more and more diagonally through the Earth's atmosphere. Our line of sight through the atmosphere is then longer and most of the blue light is scattered out before it reaches us, especially as the Sun gets very near the horizon. Relatively more red light reaches us, accounting for the reddish colour of sunsets. In other

words, the λ^{-4} dependence of the scattered intensity implies that the atmosphere scatters green, blue and violet light photons more effectively than yellow, orange, and red photons. As the Sun approaches the horizon, the path of light through the atmosphere increases, so more of the short-wavelength photons get scattered away leaving the longer-wavelength photons and the Sun look progressively redder.

4.3.3 Tyndall Scattering

Tyndall scattering is similar to Rayleigh scattering except that the condition $kR << 1$ is replaced with $kR \sim 1$ so that the wavelength is of the same order in the size of the scatterer. In this case, the scattering amplitude is obtained by evaluating the integral in equation (4.3), the scattering amplitude being given by

$$A(\theta) = 3V\gamma k^2 \frac{J_1[2kR\sin(\theta/2)]}{2kR\sin(\theta/2)}$$

where J_1 is the spherical Bessel function

$$J_1(x) = \frac{\sin(x)}{x^2} - \frac{\cos(x)}{x}.$$

In this case, the scattering is not isotropic but strongly dependent on the scattering angle. Also the intensity of the scattered field is proportional to λ^{-2} rather than λ^{-4} (under the Rayleigh scattering approximation).

4.4 The Rytov Approximation

So far in this Chapter, we have been concerned with the use of the Green function for solving two fundamental inhomogeneous partial differential equations (the Helmholtz and the Schrödinger equations). These have introduced the role that Green functions play in an important aspect of mathematical physics - scattering theory - which is fundamental to the field of image systems modelling and image understanding.

The Rytov approximation is based on the use of an exponential type or 'eikonal' transformation where a solution of the type

$$A(\mathbf{r},k)\exp[\pm s(\mathbf{r},k)] \quad \text{or} \quad A(\mathbf{r},k)\exp[\pm is(\mathbf{r},k)]$$

is considered. This is analogous (in the latter case) to a plane wave solution of the type $A\exp(\pm i\mathbf{k}\cdot\mathbf{r})$. In this transform, the scalar field s is known as the 'eikonal' from the Greek meaning 'image' or 'icon'.

The Rytov approximation is based on an idea which has a long history dating back to Huygens. In his book *A Treatise on Light*, Huygens suggested that the reflection and refraction properties of light can be explained on the basis of a sequence of wavefronts which spreads out from a source much as ripples spread out from a stone thrown into water, and that each point on such a wavefront act as a new disturbance source. Although in 1678 Huygens did not specify exactly what is meant

by a wavefront, he emphasized that the spacing between successive wavefronts need not to be uniform which is one way of considering the physical interpretation of the Rytov approximation.

4.4.1 Eikonal Transformation

Consider the 3D inhomogeneous Helmholtz equation

$$(\nabla^2 + k^2)u(\mathbf{r}, k) = -k^2\gamma(\mathbf{r})u(\mathbf{r}, k), \quad \mathbf{r} \in V.$$

If we substitute $u = u_i \exp(s)$ into this equation and differentiate, we obtain the nonlinear Riccatian equation

$$\nabla^2 s + 2\frac{\nabla u_i}{u_i} \cdot \nabla s + \nabla s \cdot \nabla s = -k^2\gamma \tag{4.5}$$

where u_i is taken to satisfy the equation

$$\nabla^2 u_i + k^2 u_i = 0, \quad \text{i.e.} \quad u_i = \exp(i\mathbf{k} \cdot \mathbf{r}).$$

Suppose we assume that s varies sufficiently slowly for the nonlinear term $\nabla s \cdot \nabla s$ in equation (4.5) to be neglected compared to the other terms, then we can write (approximately)

$$u_i \nabla^2 s + 2\nabla u_i \cdot \nabla s = -k^2 \gamma u_i. \tag{4.6}$$

This is the Rytov approximation. To facilitate a Green function solution, we substitute $s = w/u_i$ into equation (4.6). Differentiating, we have

$$\begin{gathered} u_i \nabla^2 s + 2\nabla u \cdot \nabla s \\ = \nabla^2 w + 2u_i \nabla w \cdot \nabla \left(\frac{1}{u_i}\right) + u_i w \nabla^2 \left(\frac{1}{u_i}\right) + 2\frac{\nabla u_i}{u_i} \cdot \nabla w + 2w\nabla u_i \cdot \nabla \left(\frac{1}{u_i}\right) \\ = \nabla^2 w + k^2 w \end{gathered}$$

and thus, equation (4.6) reduces to

$$\nabla^2 w + k^2 w = -k^2 \gamma u_i.$$

The Green function solution to this equation (subject to homogeneous boundary conditions) is

$$w(\mathbf{r}_0, k) = k^2 \int_V u_i(\mathbf{r}, k)\gamma(\mathbf{r})g(\mathbf{r} \mid \mathbf{r}_0, k)d^3\mathbf{r}$$

and we arrive at the solution

$$u(\mathbf{r}_0, k) = u_i(\mathbf{r}_0, k) \exp\left[\frac{k^2}{u_i(\mathbf{r}_0, k)} \int_V u_i(\mathbf{r}, k)\gamma(r)g(\mathbf{r} \mid \mathbf{r}_0, k)d^3\mathbf{r}\right].$$

We can write this result as

$$u = u_i \left(1 + \frac{k^2}{u_i} \int_V u_i \gamma g d^3\mathbf{r} + \ldots\right) \simeq u_i + k^2 \int_V u_i \gamma g d^3\mathbf{r}$$

which is the solution under the Born approximation.

4.4.2 Conditions for the Rytov Approximation

The condition required for the validity of the Rytov approximation can be investigated by considering a Green function solution with the nonlinear term $\nabla s \cdot \nabla s$ included. In this case, equation (4.6) becomes

$$u_i \nabla^2 s + 2\nabla u_i \cdot \nabla s = -k^2 \gamma u_i - u_i \nabla s \cdot \nabla s.$$

Substituting $s = w/u_i$ into this equation (except for the second term on the right hand side) we have

$$\nabla^2 w + k^2 w = -k^2 \gamma u_i - u_i \nabla s \cdot \nabla s$$

whose Green function solution is

$$w = k^2 \int_V u_i \gamma g d^3\mathbf{r} + \int_V u_i (\nabla s \cdot \nabla s) g d^3\mathbf{r}$$

so that we can write

$$s = \frac{k^2}{u_i} \int_V u_i \gamma g d^3\mathbf{r} + \frac{k^2}{u_i} \int_V u_i \gamma g \left(\frac{\nabla s \cdot \nabla s}{k^2 \gamma} \right) d^3\mathbf{r}.$$

In order for the second term on the right hand side to be neglected, we must introduce the condition

$$\frac{\nabla s \cdot \nabla s}{k^2 \gamma} << 1$$

or

$$\|k^2 \gamma\| >> \|\nabla s \cdot \nabla s\|.$$

The interpretation of this condition is not trivial. Clearly, the larger the value of k (i.e. the smaller the value of the wavelength) for a given magnitude of γ and ∇s, the more appropriate the condition becomes. Thus, the condition is valid if the wavelength of the field is small compared to γ. Since s can be taken to be the phase of the wavefield solution u, another physical interpretation of the condition is that the characteristic scale length over which a change in phase occurs ∇s is small compared to the wavelength for a given γ.

4.5 Series Solutions

The Born approximation introduced earlier was used to solve some elementary scattering problems. We shall now consider a natural extension to the Born approximation which is based on generating a series solution to the problem, known generally as the Neumann series.

4.5.1 The Born Series

Consider the 3D Green function solution to the Helmholtz equation

$$(\nabla^2 + k^2)u(\mathbf{r}, k) = -k^2\gamma u(\mathbf{r}, k)$$

which is given by

$$u(\mathbf{r}_0, k) = u_i(\mathbf{r}_0, k) + u_s(\mathbf{r}_0, k)$$

where

$$u_s(\mathbf{r}_0, k) = k^2 \int_V g(\mathbf{r} \mid \mathbf{r}_0, k)\gamma(\mathbf{r})u(\mathbf{r}, k)d^3\mathbf{r},$$

u_i is the incident field satisfying the equation

$$(\nabla^2 + k^2)u_i(\mathbf{r}, k) = 0$$

and g is the outgoing Green function given by

$$g(\mathbf{r} \mid \mathbf{r}_0, k) = \frac{\exp(ik \mid \mathbf{r} - \mathbf{r}_0 \mid)}{4\pi \mid \mathbf{r} - \mathbf{r}_0 \mid}.$$

We have seen that the Born approximation to this equation is given by considering $u \sim u_i,\ \mathbf{r} \in V$ which is valid provided $\|u_s\| << \|u_i\|$. We then obtain an approximate solution u_1, say, of the form

$$u_1(\mathbf{r}_0, k) = u_i(\mathbf{r}_0, k) + k^2 \int_V g(\mathbf{r} \mid \mathbf{r}_0, k)\gamma(\mathbf{r})u_i(\mathbf{r}, k)d^3\mathbf{r}.$$

This result can be considered to be the first approximation to a series solution, in which the second approximation u_2, say, is given by

$$u_2(\mathbf{r}_0, k) = u_i(\mathbf{r}_0, k) + k^2 \int_V g(\mathbf{r} \mid \mathbf{r}_0, k)\gamma(\mathbf{r})u_1(\mathbf{r}, k)d^3\mathbf{r}$$

and the third approximation u_3 is given by

$$u_3(\mathbf{r}_0, k) = k^2 u_i(\mathbf{r}_0, k) + \int_V g(\mathbf{r} \mid \mathbf{r}_0, k)\gamma(\mathbf{r})u_2(\mathbf{r}, k)d^3\mathbf{r}$$

and so on. In general, we can consider the iteration

$$u_{n+1}(\mathbf{r}_0, k) = u_i(\mathbf{r}_0, k) + k^2 \int_V g(\mathbf{r} \mid \mathbf{r}_0, k)\gamma(\mathbf{r})u_n(\mathbf{r}, k)d^3\mathbf{r}, \quad n = 0, 1, 2, 3, ..$$

where $u_0 = u_i$.

In principle, if this series converges, then it must converge to the solution. To investigate its convergence, it is convenient to use operator notation and write

$$u_{n+1} = u_i + \hat{I}u_n$$

where $\hat{I}$ is the integral operator

$$\hat{I} = \int_V d^3\mathbf{r} g\gamma.$$

At each iteration n we can consider the solution to be given by

$$u_n = u + \epsilon_n$$

where ϵ_n is the error associated with the solution at iteration n and u is the exact solution. A necessary condition for convergence is that $\epsilon_n \to 0$ as $n \to \infty$. Now,

$$u + \epsilon_{n+1} = u_i + \hat{I}(u + \epsilon_n) = u_i + \hat{I}u + \hat{I}\epsilon_n$$

and therefore we can write

$$\epsilon_{n+1} = \hat{I}\epsilon_n$$

since $u = u_i + \hat{I}u$. Thus

$$\epsilon_1 = \hat{I}\epsilon_0; \;\; \epsilon_2 = \hat{I}\epsilon_1 = \hat{I}(\hat{I}\epsilon_0); \;\; \epsilon_3 = \hat{I}\epsilon_2 = \hat{I}[\hat{I}(\hat{I}\epsilon_0)]; ...$$

or

$$\epsilon_n = \hat{I}^n\epsilon_0$$

from which it follows that

$$\|\epsilon_n\| = \|\hat{I}^n\epsilon_0\| \leq \|\hat{I}^n\| \times \|\epsilon_0\| \leq \|\hat{I}\|^n\|\epsilon_0\|.$$

The condition for convergence therefore becomes

$$\lim_{n\to\infty} \|\hat{I}\|^n = 0.$$

This is only possible if

$$\|\hat{I}\| < 1$$

or

$$k^2\| \int_V g(\mathbf{r} \mid \mathbf{r}_0, k)\gamma(\mathbf{r})d^3\mathbf{r}\| < 1.$$

Comparing this result with condition (4.2) and the analysis associated with it given in Section 6.2.2, it is clear that

$$\bar{\gamma} < \frac{1}{k^2R^2}$$

must be satisfied for the series to converge where R is the radius of a sphere of volume V.

This series solution, which can be written out as

$$u(\mathbf{r}_0, k) = u_i(\mathbf{r}_0, k) + k^2 \int_V g(\mathbf{r} \mid \mathbf{r}_0, k)\gamma(\mathbf{r})u_i(\mathbf{r}, k)d^3\mathbf{r} =$$

$$k^2 \int_V g(\mathbf{r} \mid \mathbf{r}_0, k)\gamma(\mathbf{r}) \left[k^2 \int_V g(\mathbf{r}_1 \mid \mathbf{r}, k)\gamma(\mathbf{r}_1)u_i(\mathbf{r}_1, k)d^3\mathbf{r}_1 \right] d^3\mathbf{r} + ...$$

$$= u_i(\mathbf{r}_0, k) + k^2 \int_V d^3\mathbf{r} g(\mathbf{r} \mid \mathbf{r}_0, k)\gamma(\mathbf{r})u_i(\mathbf{r}, k)$$

$$+k^4 \int_V \int_V d^3\mathbf{r} d^3\mathbf{r}_1 g(\mathbf{r} \mid \mathbf{r}_0, k)\gamma(\mathbf{r})g(\mathbf{r}_1 \mid \mathbf{r}, k)\gamma(\mathbf{r}_1)u_i(\mathbf{r}_1, k)$$

$$+k^6 \int_V \int_V \int_V d^3\mathbf{r} d^3\mathbf{r}_1 d^3\mathbf{r}_2 g(\mathbf{r} \mid \mathbf{r}_0, k)\gamma(\mathbf{r})g(\mathbf{r}_1 \mid \mathbf{r}, k)\gamma(\mathbf{r}_1)g(\mathbf{r}_2 \mid \mathbf{r}_1, k)\gamma(\mathbf{r}_2)u_i(\mathbf{r}_2, k)$$

$$+ \quad ...$$

is an example of a Neumann series solution to a Fredholm integral equation and is known as the Born series. The scattered field can be written in the form

$$u_s(\mathbf{r}, k) = k^2 g(r, k) \otimes \gamma(\mathbf{r})u_i(\mathbf{r}, k) + k^4 g(r, k) \otimes \gamma(\mathbf{r})[g(r) \otimes \gamma(\mathbf{r})u_i(\mathbf{r}, k)] + ...$$

where $\otimes$ denotes the three-dimensional convolution integral over V and $r \equiv \mid \mathbf{r} \mid$.

Another approach to deriving this result can be taken by considering the inverse operator. Writing

$$u = u_i + k^2 \hat{I} u$$

where

$$\hat{I} \equiv \int_V d^3\mathbf{r} \gamma(\mathbf{r}) g(\mathbf{r} \mid \mathbf{r}_0, k),$$

we have

$$(1 - k^2 \hat{I})u = u_i$$

or

$$u = (1 - k^2 \hat{I})^{-1} u_i = (1 + k^2 \hat{I} + k^4 \hat{I}^2 + k^6 \hat{I}^3 + ...)u_i.$$

Either way, the Born series can be interpreted as follows:

$$u(\mathbf{r}_0, k) = \text{incident wavefield}$$
$$+$$
$$\text{wavefield generated by single scattering events}$$
$$+$$
$$\text{wavefield generated by double scattering events}$$
$$+$$
$$\text{wavefield generated by triple scattering events}$$
$$+$$
$$\vdots$$

Each term in this series expresses the effects due to single, double, and triple, etc., scattering. Feynman diagrams can be used to represent these effects graphically,

e.g. the propagation of a wavefield generated by one interaction with another. In particle physics, interparticle interactions are complicated multiple scattering events in which the forces are transmitted by quantum fields. The propagation of fields between points is precisely what Green functions describe. So Green functions, often called Feynman propagators in particle physics, are among the standard working tools of theoretical analysis in modern quantum physics.

For an incident plane wave $u_i(\mathbf{r},k) = \exp(i\mathbf{k}\cdot\mathbf{r})$ and with $R \equiv \mid \mathbf{r}-\mathbf{r}_0 \mid$ each term in the Born series scales as $\frac{1}{R}$, $\frac{1}{R^2}$, $\frac{1}{R^3}$, etc., so that multiple-scattering gets 'weaker by the term'. This is due to the form of the Green function in 3D which scales as $1/R$, the intensity of the field being $1/R^2$ - the inverse square law. Thus, if the scattering function is characterized by a number of scattering 'sites' (i.e. where γ is composed of a distribution of point-like scatterers that are of compact support) then, provided that the distance between these sites is large, the effect of multiple scattering will be insignificant. However, if these sites are close together where the effect of the multiple scattering wavefield falling off as $1/R^2$, $1/R^3$, etc., is not appreciable, then multiple scattering events will contribute significantly to the scattered field. Hence, one way to interpret the meaning of 'weak' and 'strong' scattering is in terms of the 'density' of scattering sites over the volume V being low or high, respectively. For $\lambda \sim R$ where R is the characteristics size of the scatterer, the Born approximation holds provided the root mean square of the scattering function over the volume is much less than 1. This is a quantification of the principle that the density of scattering sites from which we can suppose the scattering function is composed is low.

Another important feature of the Born series for Helmholtz scattering is that the terms are scaled by k^2, k^4, k^6. Thus for a fixed $k << 1$ (long wavelength waves),

$$u(\mathbf{r}_0,k) = u_i(\mathbf{r}_0,k) + k^2 \int_V g(\mathbf{r} \mid \mathbf{r}_0,k)\gamma(\mathbf{r})u_i(\mathbf{r},k)d^3\mathbf{r}.$$

In 1D, the Green function scales as $1/k$, the Born series for Helmholtz scattering being given by

$$u(x_0,k) = u_i(x_0,k) + \frac{ik}{2}\int_L dx \exp(ik \mid x-x_0 \mid)\gamma(x)u_i(x,k)$$

$$-\frac{k^2}{4}\int_L\int_L dxdx_1 \exp(ik \mid x-x_0 \mid)\gamma(x)\exp(ik \mid x_1-x \mid)\gamma(x_1)u_i(x_1,k)$$

$$-\frac{ik^3}{8}\int_L\int_L\int_L dxdx_1dx_2 \exp(ik \mid x-x_0)\gamma(x)\exp(ik \mid x_1-x \mid)\gamma(x_1)$$

$$.\exp(ik \mid x_2-x_1)\gamma(x_2)u_i(x_2,k)$$

$$+ \ ...$$

In this case, the series does not get 'weaker by the term' according to $1/R^n$ but by $1/2^n$. Consequently, we should expect that multiple scattering is a more common

occurrence when waves scatter (transmit/reflect) from layered materials. This is readily experienced when observing light reflecting from two glass plates - double glazing for example. Here, a number of faded 'ghost images' are seen in addition to the two primary images obtained from the partial reflection of light by the first plate and that from the second. As in the 3D case, the Born approximation 'improves' at larger wavelengths since for $k << 1$

$$u(x_0,k) = u_i(x_0,k) + \frac{ik}{2}\int_L \exp(ik \mid x - x_0 \mid)\gamma(x)u_i(x,k)dx.$$

In quantum (Schrödinger) scattering, the Born series is of the same form but without the factors of k^2, k^4, k^6, etc., and in 1D is given by

$$u(x_0,k) = u_i(x_0,k) + \int dx g(x \mid x_0,k)\gamma(x)u_i(x,k)dx + ...$$

Now, the 1D Green function is given by

$$g(x \mid x_0,k) = \frac{i}{2k}\exp(ik \mid x - x_0 \mid)$$

and so for $k >> 1$

$$u(x_0,k) = u_i(x_0,k) + \frac{i}{2k}\int \exp(ik \mid x - x_0 \mid)V(x)\exp(ikx)dx$$

$$= u_i(x_0,k) + \exp(-ikx_0)\frac{i}{2k}\int V(x)\exp(2ikx)dx, \quad x_0 \to \infty.$$

Thus, for very high frequency quantum wavefields in 1D, the Fourier transform of the scattering potential γ is an exact scattering transform. This result can be applied to the 1D inhomogeneous Helmholtz equation by mapping it into the Schrödinger equation. Writing

$$\left(\frac{\partial^2}{\partial x^2} + k^2\Gamma(x)\right)u(x,k) = 0$$

where

$$\Gamma(x) = 1 + \gamma(x),$$

application of the Liouville transformation

$$U(y,k) = g(x)u(x,k), \quad \frac{dx}{dy} = \frac{1}{[g(x)]^2}, \quad \text{and} \quad g(x) = \Gamma^{\frac{1}{2}}(x)$$

gives

$$\left(\frac{\partial^2}{\partial y^2} + k^2\right)U(y,k) = f(y)U(y,k)$$

where

$$f(y) = \frac{1}{g(y)}\frac{\partial^2}{\partial y^2}g(y).$$

In imaging science, the fundamental imaging equation comes from assuming that the recorded data $d = u - u_i$ is of the form

$$d = k^2 \hat{I} u_i + n$$

where

$$n = (k^4 \hat{I}^2 + k^6 \hat{I}^3 + ...) u_i + \text{other noise}$$

In other words, the multiple scattering events are assumed to be part of the noise inherent in the system recording.

One of the principal issues with using the Born approximation is that it is generally going to be valid for the case when $\lambda >> R$ where R is a measure of the characteristic size of the inhomogeneity. However, in imaging science, to obtain information on an object over a scale of R, we apply wavefields whose wavelength is of the same order, i.e. $\lambda \sim R$. Now, since the Born approximation requires that (in 3D)

$$\bar{\gamma} << \frac{1}{kR},$$

for $\lambda \sim R$,

$$\bar{\gamma} << 1.$$

In other words, to utilize the fundamental imaging equation (which is a product of applying the Born approximation) the material we are imaging should ideally have inhomogeneities whose root mean square value is much less than 1. Because this condition is not always satisfied, multiple scattering effects are inevitable. Nevertheless, the basic imaging model is, for better or worse, based on the Born scattering term plus noise for the case when $\lambda \sim R$. Asymptotic conditions such as $\lambda \to \infty$ or $\lambda \to 0$ may provide exact scattering solutions but they are inconsistent with imaging systems based on the use of radiation where $\lambda \sim R$.

When multiple scattering is a dominant feature of an image system, although it may be possible to construct a deterministic multiple scattering model, the application of such a model for the development of a practical image reconstruction and image processing algorithms is often intractable. Instead we can consider the wavefield generated by multiple scattering events to be a stochastic field and investigate its characteristics using statistical modelling and analysis. This approach is of course consistent with many areas of physics and engineering when the 'physics' that one is attempting to model becomes too complicated for a deterministic analysis to be of any practical value. In such cases we turn to statistical methods of modelling the data.

4.5.2 The Rytov Series

It is worth mentioning that a Rytov series can be derived by extending the Rytov approximation in the same way that the Born series has been derived here by extending the Born approximation and considering higher order iterates subject to a condition for convergence being satisfied. However, the interpretation of the Rytov series is not trivial and the computational effort required to evaluate the series for a given scattering function can become problematic.

The use of the Born series (and the Rytov series) leads to computational problems when evaluating a fully multiple scattered field. First is the issue over the convergence criterion for the series which may not always be satisfied; second is the issue of the singularities that arise when a multiple point scattering model (i.e. multiple delta functions located at different position in space) for the scattering function is introduced into the Born series. These problems necessitated the development of renormalization theory in the early 1960s which lies beyond the scope of this work. However, it is worth noting, that issues concerning the development of renormalization theory and the difficulties associated with its application helped to forge the foundations of string theory that, to this day, remains the most promising approach for the development of a unified field theory (a theory of everything!).

4.6 Inverse Scattering

Inverse scattering aims to reconstruct the scattering function from measurements of the data. The practicability of solving inverse scattering problems analytically and implementing them experimentally varies considerably from one application to another. An inversion method is usually based on the approximation that has been applied to solve the forward scattering problem given the wave equation. For example, given the Helmholtz equation, then under the Born approximation, in the far field region, the scattering amplitude is given by

$$A(\hat{\mathbf{n}}_s, \hat{\mathbf{n}}_i, k) = k^2 \int_V \exp[-ik(\hat{\mathbf{n}}_s - \hat{\mathbf{n}}_i) \cdot \mathbf{r}]\gamma(\mathbf{r}) d^3\mathbf{r}.$$

The inverse solution to this problem is therefore compounded in the inverse Fourier transform. In 1D, the solution is, for a unit plane wave,

$$u(x_0, k) = \exp(ikx_0) + \exp(-ikx_0)r(k)$$

where r is the 'reflection coefficient' given by

$$r(k) = \frac{ik}{2} \int_L \gamma(x) \exp(2ikx) dx$$

which can be written as

$$r(k) = \frac{1}{4} \int dx \exp(ikx) \frac{d}{dx} \gamma(x/2).$$

Hence, inversion is achieved by taking the inverse Fourier transform and integrating the result.

The link between the application of the Born approximation in the far field and the Fourier transform should now be clear. This 'link' is essential in imaging science and is why the Fourier transform plays such an essential role. Inverse solutions under the Born approximation are in effect the same as implementing Fourier based

reconstruction methods in imaging science, at least when the data collected are the result of a scattering event. In some cases, the scattering is not as weak as it should be to support application of the Born approximation. In such cases, Fourier based image reconstructions can become distorted. There is, however, a method of inverting a wavefield that is the result of multiple Born scattering; this is known as the Jost-Kohn method first published in 1952. A brief overview of this method follows.

Using operator notation, the Born series can be written as

$$u = u_i + \hat{I}_i\gamma + \hat{I}_i(\gamma\hat{I}\gamma) + I_i[\gamma\hat{I}(\gamma\hat{I}\gamma)] + ...$$

where γ is either the scattering potential (for Schrödinger scattering) or $k^2\gamma$ (for Helmholtz scattering) and

$$\hat{I}_i = \int d^3\mathbf{r}u_i g, \quad \hat{I} = \int d^3\mathbf{r}g.$$

Now, let $\epsilon U = u - u_i$ and

$$\gamma = \sum_{j=1}^{\infty} \epsilon^j \gamma_j.$$

Then

$$\epsilon U = \hat{I}_i[\epsilon\gamma_1 + \epsilon^2\gamma_2 + \epsilon^3\gamma_3 + ...]$$

$$+\hat{I}_i[(\epsilon\gamma_1 + \epsilon^2\gamma_2 + \epsilon^3\gamma_3 + ...)\hat{I}(\epsilon\gamma_1 + \epsilon^2\gamma_2 + \epsilon^3\gamma_3 + ...)]$$

$$+\hat{I}_i\{(\epsilon\gamma_1 + \epsilon^2\gamma_2 + \epsilon^3\gamma_3 + ...)\hat{I}[(\epsilon\gamma_1 + \epsilon^2\gamma_2 + \epsilon^3\gamma_3 + ...)$$

$$\hat{I}(\epsilon\gamma_1 + \epsilon^2\gamma_2 + \epsilon^3\gamma_3 + ...)]\} + ...$$

Equating terms with common coefficients ϵ, ϵ^2, etc., we have

For $j = 1$:

$$U = \hat{I}_i\gamma_1; \quad \gamma_1 = \hat{I}_i^{-1}U.$$

For $j = 2$:

$$0 = \hat{I}_i\gamma_2 + \hat{I}_i(\gamma_1\hat{I}\gamma_1); \quad \gamma_2 = -\hat{I}_i^{-1}[\hat{I}_i(\gamma_1\hat{I}\gamma_1)]$$

and so on. By computing the functions γ_j using this iterative method, the scattering function γ is obtained by summing γ_j for $\epsilon = 1$. This approach provides a formal exact inverse scattering solution but it is not unconditional, i.e. the inverse solution is only applicable when the Born series converges to the exact scattering solution and thus when

$$\| \int_V g(\mathbf{r} \mid \mathbf{r}_0, k)\gamma(\mathbf{r})d^3\mathbf{r}\| < 1$$

We note, that, for $j = 1$, the solution for γ_1 is that obtained under the Born approximation.

4.7 Exact Inverse Scattering Solution

For $\gamma(\mathbf{r}) \to 0$ as $r \equiv \mid \mathbf{r} \mid \to \infty$, the the Lippmann-Schwinger equation [14]

$$u(\mathbf{r},k) = u_i^{\pm}(\mathbf{r},k) + k^2 g(r,k) \otimes_3 \gamma(\mathbf{r}) u\mathbf{r},k)$$

where $\otimes_3$ denotes the three-dimensional convolution integral and $u_i^{\pm} = \exp(\pm ik\hat{\mathbf{n}}_i \cdot \mathbf{r})$ is a solution of

$$(\nabla^2 + k^2)u(\mathbf{r},k) = 0)$$

This is the solution of the inhomogenous Helmholtz equation

$$(\nabla^2 + k^2)u(\mathbf{r},k) = -k^2\gamma(\mathbf{r})u(\mathbf{r},k)$$

Instead of computing u given γ before tackling the inverse scattering problem, we now consider an inverse scattering solution that is based on the transformation of the inhomogeneous Helmholtz equation to the form

$$-k^2\gamma(\mathbf{r}) = \frac{u^*(\mathbf{r},k)}{\mid u(\mathbf{r},k) \mid^2} \nabla^2 \left(u_s(\mathbf{r},k) - \frac{k^2}{4\pi r} \otimes_3 u_s(\mathbf{r},k) \right) \tag{4.7}$$

A derivation of this transformation is given in Appendix 1.

Since

$$\|u_s - (k^2/4\pi r) \otimes_3 u_s\|_2 \leq \|u_s\|_2[1 + k^2\sqrt{r/(4\pi)}],$$

in the far field, equation (4.7) becomes

$$-k^2\gamma = \frac{-1}{u_i^{\pm} + u_s} k^2 u_s \otimes_3 \nabla^2 \left(\frac{1}{4\pi r} \right) = k^2 A^{-1}[(u_i^{\pm})^* + u_s^*]u_s, \quad r \to \infty \tag{4.8}$$

where $A^{-1} = \mid u_i^{\pm} + u_s \mid^{-2}$. Fourier analysis of equation (4.8) provides a far-field solution for the scattered field that is compatible with the result under the Born approximation, i.e. Fourier-space far-field equivalence. Taking the Fourier transform of equation (4.8) and using the product theorem, for u_i^-, we obtain

$$\widetilde{\gamma}(k\hat{\mathbf{n}}) = [\widetilde{u}_s[k(\hat{\mathbf{n}}_i - \hat{\mathbf{n}})] + \widetilde{u}_s^*(k\hat{\mathbf{n}}) \otimes_3 \widetilde{u}_s(k\hat{\mathbf{n}})] \otimes_3 \widetilde{A}^{-1}(k\hat{\mathbf{n}}) \tag{4.9}$$

where $\hat{\mathbf{n}} = \mathbf{k}/k$ and, critical to the arguement, $\widetilde{u}_s$ is taken to be u_s in the far field. Since $A^{-1} = 1 - u_i^- u_s^* - u_s(u_i^-)^* - \mid u_s \mid^2 + ...$,

$$\widetilde{A}^{-1}(k\hat{\mathbf{n}}) = \delta^3(k\hat{\mathbf{n}}) - \widetilde{u}_s^*[k(\hat{\mathbf{n}}_i + \hat{\mathbf{n}})] - \widetilde{u}_s[k(\hat{\mathbf{n}}_i - \hat{\mathbf{n}})] - \widetilde{u}_s(k\hat{\mathbf{n}}) \otimes_3 \widetilde{u}_s^*(k\hat{\mathbf{n}}) + ...$$

With $\hat{\mathbf{n}}_i - \hat{\mathbf{n}} = \hat{\mathbf{n}}_s$ equation (4.9) can be written in the form

$$\widetilde{u}_s(k\hat{\mathbf{n}}_s) \otimes_3 \widetilde{A}^{-1}[k(\hat{\mathbf{n}}_i - \hat{\mathbf{n}}_s)] = \widetilde{\gamma}[k(\hat{\mathbf{n}}_i - \hat{\mathbf{n}}_s)]$$

$$-\widetilde{u}_s^*[k(\hat{\mathbf{n}}_i - \hat{\mathbf{n}}_s)] \otimes_3 \widetilde{u}_s[k(\hat{\mathbf{n}}_i - \hat{\mathbf{n}}_s)] \otimes_3 \widetilde{A}^{-1}[k(\hat{\mathbf{n}}_i - \hat{\mathbf{n}}_s)]$$

Note that for back-scattering (when $\hat{\mathbf{n}}_i = -\hat{\mathbf{n}}_s$) equation (4.9) becomes

$$\widetilde{\gamma}(k\hat{\mathbf{n}}_s) = [\widetilde{u}_s(-2k\hat{\mathbf{n}}_s) + \widetilde{u}_s^*(k\hat{\mathbf{n}}_s) \otimes_3 \widetilde{u}_s(k\hat{\mathbf{n}}_s)] \otimes_3 \widetilde{A}^{-1}(k\hat{\mathbf{n}}_s)$$

The scattering function is obtained directly from data on the far-scattered-field based on equation (4.9).

4.8 Computation of the Scattered Field

Unlike the inverse scattering solution, computation of the unconditional scattered field requires, like the Born or Rytov series, an iterative procedure, i.e.

$$\widetilde{u}_s^{m+1}(k\hat{\mathbf{n}}_s) = \widetilde{\gamma}[k(\hat{\mathbf{n}}_i - \hat{\mathbf{n}}_s)] \otimes_3 \widetilde{A}^m[k(\hat{\mathbf{n}}_i - \hat{\mathbf{n}}_s)] - \widetilde{u}_s^m[k(\hat{\mathbf{n}}_i - \hat{\mathbf{n}}_s)] \otimes_3 (\widetilde{u}_s^m)^*[k(\hat{\mathbf{n}}_i - \hat{\mathbf{n}}_s)] \tag{4.10}$$

where

$$\widetilde{u}_s^0(k\hat{\mathbf{n}}_s) = \widetilde{\gamma}[k(\hat{\mathbf{n}}_i - \hat{\mathbf{n}}_s)]$$

the back-scattered cross section being obtained by computing $\mid \widetilde{u}_s^m(k\hat{\mathbf{n}}_s) \mid^2$, $\hat{\mathbf{n}}_s \sim -\hat{\mathbf{n}}_i$.

4.8.1 Approximation for $\widetilde{A}^{-1} = \delta^3$

It is clear, that both equations (4.9) and (4.10) can be significantly simplified if we assume that $\mid u_i^{\pm} + u_s \mid^2 \sim 1$ because in this case $\widetilde{A} = \delta^3$ and $\widetilde{A}^{-1} = \delta^3$ so that these equation can be approximated by the results

$$\widetilde{\gamma}(k\hat{\mathbf{n}}) = \widetilde{u}_s[k(\hat{\mathbf{n}}_i - \hat{\mathbf{n}})] + \widetilde{u}_s^*(k\hat{\mathbf{n}}) \otimes_3 \widetilde{u}_s(k\hat{\mathbf{n}})$$

and

$$\widetilde{u}_s^{m+1}(k\hat{\mathbf{n}}_s) = \widetilde{\gamma}[k(\hat{\mathbf{n}}_i - \hat{\mathbf{n}}_s)] - \widetilde{u}_s^m[k(\hat{\mathbf{n}}_i - \hat{\mathbf{n}}_s)] \otimes_3 (\widetilde{u}_s^m)^*[k(\hat{\mathbf{n}}_i - \hat{\mathbf{n}}_s)]$$

respectively. It is then apparent that multiple scattering effects are compounded in a single term, namely, the term $\widetilde{u}_s^m[k(\hat{\mathbf{n}}_i - \hat{\mathbf{n}}_s)] \otimes_3 (\widetilde{u}_s^m)^*[k(\hat{\mathbf{n}}_i - \hat{\mathbf{n}}_s)]$. The effect of using this approximation is explored further in Chapter 6 which includes numerical results on a study for the one-dimensional case and is used to develop a model for a side-band pulse-echo signal that is inclusive of multiple scattering under the condition that $\widetilde{A}^{-1} \sim \delta^3$. However, it is clear that underder this condition, the Born approximation can now be attributed to the case when $\widetilde{u}_s^m[k(\hat{\mathbf{n}}_i - \hat{\mathbf{n}}_s)] \otimes_3 (\widetilde{u}_s^m)^*[k(\hat{\mathbf{n}}_i - \hat{\mathbf{n}}_s)] \sim 0$ which 'translates' to the autoconvolution of the scattered field being effectively zero. The physical interpretation of this result is that multiple scattering processes can be expected to produce replicating patterns in the scattered field. These 'matching features' will then contribute to the autoconvolution function making it a non-zero function.

4.8.2 Approximation under the Skew Hermitian Condition

Another simplification can be made to equations (4.9) and (4.10) is the product of the incident and scattered fields are taken to be Skew Hermitian, i.e.

$$u_i^- u_s^* = -(u_i^- u_s^*)^*$$

so that $A = 1 + \mid u_s \mid^2$ and $A^{-1} = 1 - \mid u_s \mid^2 + \ldots$. In particular, equation (4.10) becomes

$$\widetilde{u}_s^{m+1}(k\hat{\mathbf{n}}_s) = \widetilde{\gamma}[k(\hat{\mathbf{n}}_i - \hat{\mathbf{n}}_s)] + \widetilde{\gamma}[k(\hat{\mathbf{n}}_i - \hat{\mathbf{n}}_s)] \otimes_3 \widetilde{u}_s^m[k(\hat{\mathbf{n}}_i - \hat{\mathbf{n}}_s)] \otimes_3 (\widetilde{u}_s^m)^*[k(\hat{\mathbf{n}}_i - \hat{\mathbf{n}}_s)]$$

$$-\widetilde{u}_s^m[k(\hat{\mathbf{n}}_i - \hat{\mathbf{n}}_s)] \otimes_3 (\widetilde{u}_s^m)^*[k(\hat{\mathbf{n}}_i - \hat{\mathbf{n}}_s)] \tag{4.11}$$

which includes the convolution of the Born approximated scatterd field with the autoconvolution of the field (on an iterative basis). If we defined the three-dimensional Fourier scattering operator $\mathcal{F}_3$ as

$$\mathcal{F}_3 \equiv \int_{-\infty}^{\infty} d^3\mathbf{r} \exp[-ik(\hat{\mathbf{n}}_i - \hat{\mathbf{n}}_s) \cdot \mathbf{r}]$$

then, for $m = 1$, we can write equation (4.11) in the form

$$\widetilde{u}_s^1(k\hat{\mathbf{n}}_s) = \mathcal{F}_3[\gamma(\mathbf{r})] - \mathcal{F}_3(\mid \gamma(\mathbf{r}) \mid^2 [1 - \gamma(\mathbf{r})]) \tag{4.12}$$

and it is now clear that if $\|\gamma(\mathbf{r})\| << 1$, then

$$\widetilde{u}_s^1(k\hat{\mathbf{n}}_s) \sim \widetilde{u}_s^0(k\hat{\mathbf{n}}_s) = \mathcal{F}_3[\gamma(\mathbf{r})]$$

4.8.3 Scattering from a Radially Symmetric Dielectric

A comparison is made of the differences in the scattered field generated by the zero order $\widetilde{u}_s^0$ and first order $\widetilde{u}_s^1$ terms in equation (4.12) for a radially symmetric scatterer. For a radially symmetric scatterer $\gamma(\mathbf{r}) = \gamma(r)$, the Fourier scattering operator is given by

$$\mathcal{F}_3 \equiv \frac{4\pi}{\xi} \int_0^{\infty} dr \sin(\xi r)$$

operating on a mdified scattering function given by $\gamma_r(r) = r\gamma(r)$ and where $\xi = 2k\sin(\theta/2)$, θ being the scattering angle over 2π radians. The integral is a sine transform, i.e.

$$\widetilde{\gamma}_r(\xi) = \mathcal{S}[\gamma_r(r)] = \int_0^{\infty} \gamma_r(r) \sin(\xi r) dr$$

This transform can be computed using a Fourier transform via application of the Hilbert transform because the Hilbert transform provides a single sided representation of the spectrum of a function. Thus we can write

$$\widetilde{\gamma}_r(\xi) = \mathrm{Im}\{\mathcal{H}\mathcal{F}[\gamma_r(r)]\}$$

where $\mathcal{H}$ denotes the Hilbert transform given by

$$\mathcal{H}f(r) = -\frac{1}{\pi r} \otimes f(r)$$

with spectral response

$$-\frac{1}{\pi r} \otimes f(r) \leftrightarrow \widetilde{s}(k)\widetilde{f}(k)$$

where $\widetilde{f}(k)$ is the Fourier transform of $f(r)$, $\widetilde{s}$ is the Heaviside step function given by

$$\widetilde{s}(k) = \begin{cases} 1 & \forall k \geq 0; \\ 0 & \forall k < 0. \end{cases}$$

and we note that

$$\widetilde{s}(k) \leftrightarrow \delta(r) + \frac{i}{\pi r}$$

Using this approach, from equation (4.12), for constant k the zero and first order scattered field can be computed using the following equations respectively (ignoring scaling by 4π):

$$\widetilde{u}_s^0(\theta) = \frac{1}{2k\sin(\theta/2)}\mathrm{Im}\{\mathcal{HF}[r\gamma(r)]\}[2k\sin(\theta/2)] \tag{4.13}$$

and

$$\widetilde{u}_s^1(\theta) = \widetilde{u}_s^0(\theta) - \frac{1}{2k\sin(\theta/2)}\mathrm{Im}\{\mathcal{HF}(r \mid \gamma(r) \mid^2 [1-\gamma(r)])\}[2k\sin(\theta/2)] \tag{4.14}$$

Figures 4 to 6 provide an example comparative study of the scattered field computed using equations (4.13) and (4.14) for three different dielectric scattering functions $\gamma(r)$ whose amplitude has been chosen to minimise the effect of second and higher order scattering effects, i.e. where the numerical value of $\gamma^n, n > 3$ is insignificant. In Figure 4 the differences in the scattered fields are not particularly significant because the scattering function is a homogenous dielectric sphere where multiple scattering are not significant (other than those generated from internal 'boundary scattering'). However, when the sphere is a layered dielectric with well defined and sharp discontinuities in the value of γ, differences are apparent as illustrated in Figure 5 particularly with regard to the relative amplitudes of the side-lobes. For a randomly distributed layered dielectric sphere, Figure 6, there is a noticeable difference between the zero order and first order back-scattered fields.

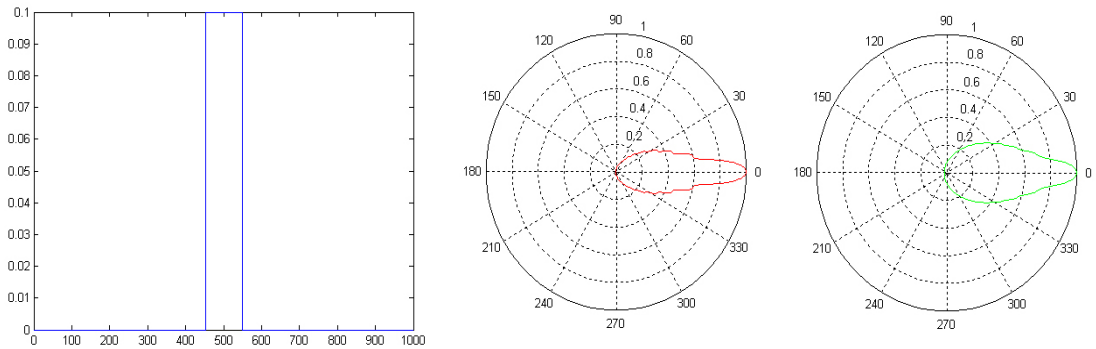

FIGURE 4 Log-polar plots for the zero (center) and first order (right) scattered field based on the Skew Hermitian condition generated by a uniformly distributed (radially symmetric) dielectric scattering function $\gamma(r)$ of compact support (left).

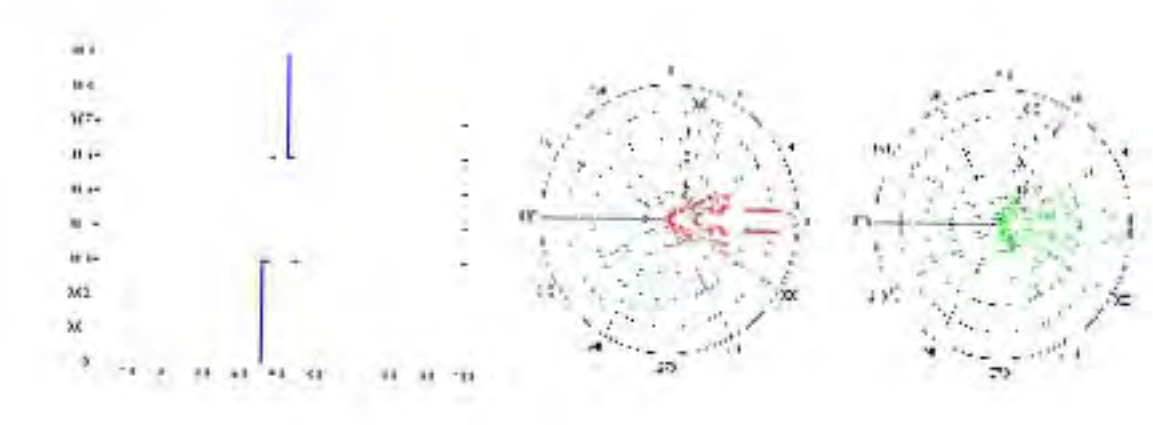

FIGURE 5 Log-polar plots for the zero (center) and first order (right) scattered fields based on the Skew Hermitian condition generated by a non-uniformly distributed (radially symmetric) dielectric scattering function $\gamma(r)$of compact support (left).

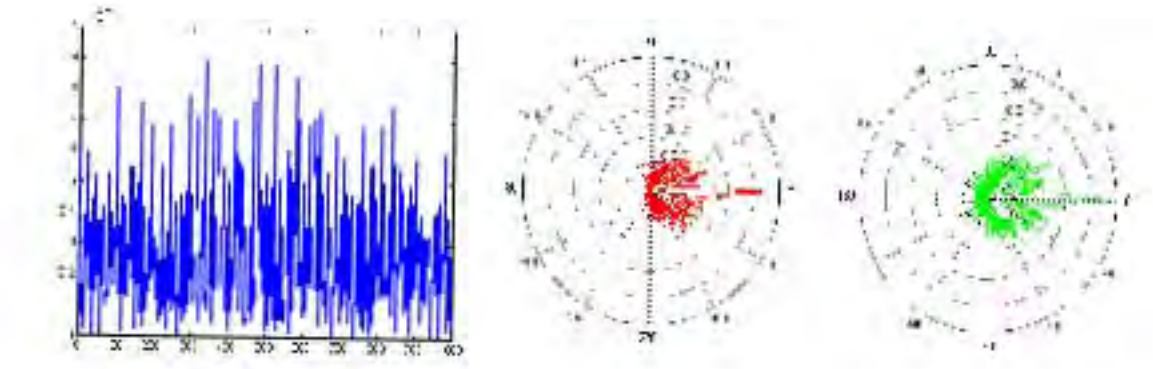

FIGURE 6 Log-polar plots for the zero (center) and first order (right) scattered fields based on the Skew Hermitian condition generated by a Gaussian distributed (radially symmetric) random dielectric scattering function $\gamma(r)$ (left).

4.9 Discussion

The purpose of this chapter has been to review formal methods in scattering theory for modelling and computing the scattered field, particularly with regard to those methods that are used in image modelling and imaging systems analysis. Although the inverse solution presented in Section 4.6 for the Born series is nearly fifty years old, it has not found any practically useful applications in imaging science and nothing has been found in the open literature in this regard. Inverse scattering solutions in electromagnetism are designed and implemented under the Born approximation and is some cases, the Rytov approximation. The fundamental imaging equation, which provides a mapping between the object and image planes in terms of a convolution integral, is an intrinsically product of using the Born approximation, irrespective of the application. Note that this result can be derived using surface scattering theory by applying the Kirchhoff approximation which is often used in the theory of optical image analysis, for example [18]. The Kirchhoff approximation is the surface scattering equivalent of the Born approximation.

Section 4.7 has introduced a 'direct' approach to solving the inverse scattering problem. In principle, this provides an exact inverse scattering solution that is not based on an 'iterative solution to an iterative solution' (i.e. the Jost-Kohn method given in Section 4.6). However, computation of the scattered field does require iteration which has been briefly investigated in Section 4.8.

The principle associated with with the method adopted is based on equation (4.8) in which we have taken the Fourier transform of the equation

$$\gamma = A^{-1}[(u_i^{\pm})^* + u_s^*]u_s \tag{4.15}$$

where the Fourier transform of u_s is taken to be the scattered field. This is based on utilising the result obtained under the Born approximation in the far field to specify the transform that is taken. However, in general, one can consider any transform of this equation where the transform of u_s is taken to be the scattered field generated under a given regime. Thus, in the intermediate field, for example, when the scattered field is measured in the Fresnel zone, the Fresnel transform (obtained by applying the Fresnel approximation to the Green's function as discussed in Chapter 3) of equation (4.15) is used to develop a scattering model.

In the following chapter, the Born approximation is used to model a Synthetic Aperture Radar (SAR) imaging system which represents a case study on the applications of the weak scattering approximation. In Chapter 6, the exact inverse scattering method discussed in this Section 4.7 is developed further under the condition discussed in Section 4.8.1 and the result used to establish a 'filtering protocol' for SAR images.

5 AN ELECTROMAGNETIC SCATTERING MODEL FOR SAR

Radar (**Ra**dio **d**etection **a**nd **r**anging) has been used for many years to detect airborne objects using ground based antennas and to image the 'ground truth' using airborne platforms. The world's first ever Radar system was constructed in Britain in the late 1930s. It was originally based on using CW radio wavefields. When these radio waves were reflected from an object, a modulation in the amplitude of the return signal occurred providing a characteristic detection signature. The resolution of this Radar system was very poor due to the long wavelength ($\sim$ 1km) radio waves that were available at the time but it was instrumental in tracking enemy aircraft and giving estimates of their direction and number during the 'Battle of Britain' in the late summer of 1940.

Research undertaken at Birmingham University in the early 1940s led to the development of the cavity magnetron. In a strong magnetic field electrons gyrate around the direction of the field-lines at a high frequency to produce radio waves with a much shorter wavelength ($\lambda \sim 10^{-3}$km). These are known today as microwaves. This technology was used almost immediately for navigation in the night bomber offensive of 1943-45. Microwave pulses were used to generate an image of the ground-truth by rotating the antenna (a microwave 'horn'). Major advances in microwave (Radar) technology occurred in Britain and Germany throughout the early 1940s, and a new research and development laboratory was established at the Massachusetts Institute of Technology, USA, to advance the systems provided to the Americans by the British as part of the lend-lease policy. The technology at the time was based on using sideband pulses. The range resolution was determined by the width of the pulse, and the lateral or azimuth resolution by the width of the beam at the range required. This was the basis for most of the Radar systems used up until the early 1960s when an American invention led to a radical improvement in the range resolution. This was achieved by linearly frequency modulating the pulse and then matched filtering the return 'echo' with its complex conjugate. The frequency modulation was achieved by linearly increasing the intensity of the magnetic field in the cavity magnetron over the duration of a pulse. Further developments in the 1960s and early 1970s paved the way for a new generation of high resolution

Radar systems which helped lead to the development of Synthetic Aperture Radar in the mid 1970s (although it had been used covertly for military and some space programmes well before that time). Synthetic Aperture Radar (SAR) was developed to study the surface of the Earth (and other planets) from both spaceborne and airborne platforms [19], [20]. The basic difference between spaceborne and airborne SAR is the 'look-down' angle of the microwave beam that is used. Spaceborne SAR uses look-down angles of $\sim 70^0$ whereas airborne systems use look-down angles $\sim 10^0$. Both systems attempt to classify the inhomogeneous nature of the Earth's surface by repeatedly emitting a pulse of microwave radiation and recording the back-scattered field [21], [22]

In this Chapter attention is focused on airborne SAR which is now used extensively for both civilian and military reconnaissance. The original spaceborne SAR - Seasat - became operational soon after its launch in June 1978 but only functioned for a limited period of time (from July to October of the same year) owing to malfunction. It was designed to carry out studies of the ocean surface using a range of microwave sensors and was equipped with a 24 cm wavelength SAR. Another satellite system - Earth Resources Satellite (ERS-1) was launched in the early 1990s and included a 5cm wavelength SAR.

A conventional side-looking radar (a real aperture radar) operating at many tens of kilometres is only able to obtain lateral resolutions of about a kilometre. By synthesizing the aperture of the radar, one can obtain resolutions of a few metres. This enhancement of resolution by three orders of magnitude, together with the fact that radar can be used in cloud or fog, means that SAR is ideal for airborne reconnaissance. The quantity of data that must be recorded and processed is typically a million independent pixels (discrete picture elements) every second. This immense amount of data has to be examined and positions of interest (targets) identified and extracted, ideally, in near real time.

Another important aspect of SAR reconnaissance is that, in comparison with optical or infrared reconnaissance, radar can often pick out details on the ground which are either invisible or difficult to distinguish with the human eye. For example, it is possible to distinguish between different types of vegetation. In some cases it is even possible to observe sub-surface structures in regions where the skin depth of the ground is small and the radar can penetrate a short distance into the ground. Many ground-based objects are good reflectors of microwave radiation, particularly those objects that are composed from materials that are good conductors (i.e. metallic objects which have a relatively large radar cross-section). Objects of this kind can therefore be distinguished more easily using radar reconnaissance. This is why airborne SAR imaging is often used for the surveillance of military hardware.

SAR systems are usually classified in terms of the wavelength that is used. The two basic modes of operation are X-band, with a wavelength of 2.8 cm, and L-band, with a wavelength of 24 cm. In addition to different wavelengths, different polarizations can be used. One of the most commonly used types is vertical polarization. This is where an electric field is emitted which points in the vertical direction (relative to the orientation of the antenna). The back-scattered field that is produced with the same polarization is then measured. For this reason, the type

of data produced is called vertical-vertical or VV polarization data. In addition to the vertically polarized return, scattering by the ground creates polarizations which differ from that of the incident electric vector, one component being along the direction of the horizontal axis. This is known as the depolarized return, and the type of data that is produced by measuring are known as vertical-horizontal or VH polarization data. Alternatively, an incident electric field can be produced where the electric vector points along the horizontal axis. The data produced by measuring the like polarized field is known as the horizontal-horizontal or HH polarization data. The data produced by measuring the cross-polarized return in this case is known as the horizontal-vertical or HV polarization data. Hence, in principle, there are four modes of operation that can be used. In practice VH and HV SAR images are not significantly different. However, the difference between VV, HH SAR images can be considerable.

An example of a SAR 'microwave image' together with an incoherent optical image of approximately the same region is given in Figure 7. These images are of a region of Northamptonshire (just south-west of the town of Northampton), England, and show both urban (e.g. the village of Wootton) and rural features. The major road in the bottom left hand corner of this image is the M1 Motorway (which runs from London to Leeds, Yorkshire) in the locality of Junction 15. In contrast to the optical image, the SAR image is dominated by noise of a special and quantifiable physical type, namely, speckle.

The SAR image given in Figure 7 is an airborne SAR is VV polarized. This type of image is known as a VVX SAR image (VV for vertical-vertical polarization and X for X-band). Each resolution cell in this image corresponds to a real length of about 1.5m. The image was obtained at a range of approximately 50 km and an altitude of about 8 km. There are a number of interesting features in this SAR image. A close inspection reveals that there is a variety of textures which change from one region of the image to the next. These textures are related to physical changes in the terrain such as the type of vegetation that is present. There is a particularly marked difference between rural and urban regions. majority of these buildings being constructed from non-conductive materials (brick, concrete and wood, etc.).

In general, much stronger reflections occur from structures that are made from conductive materials.

5.1 Principles of SAR

Synthetic aperture radar is a pulse-echo system which utilizes the response of a scatterer as it passes through the beam to synthesize the lateral (azimuth) resolution. This allows relatively high resolution images to be obtained at a long range. The basic geometry of the system is given in Figure 8. Here, and throughout the rest of this chapter, the range coordinate is denoted by x and the tracking coordinate along the flight path is denoted by y. The latter coordinate is referred to as the azimuth direction. The antenna emits a pulse of microwave radiation and the return signal

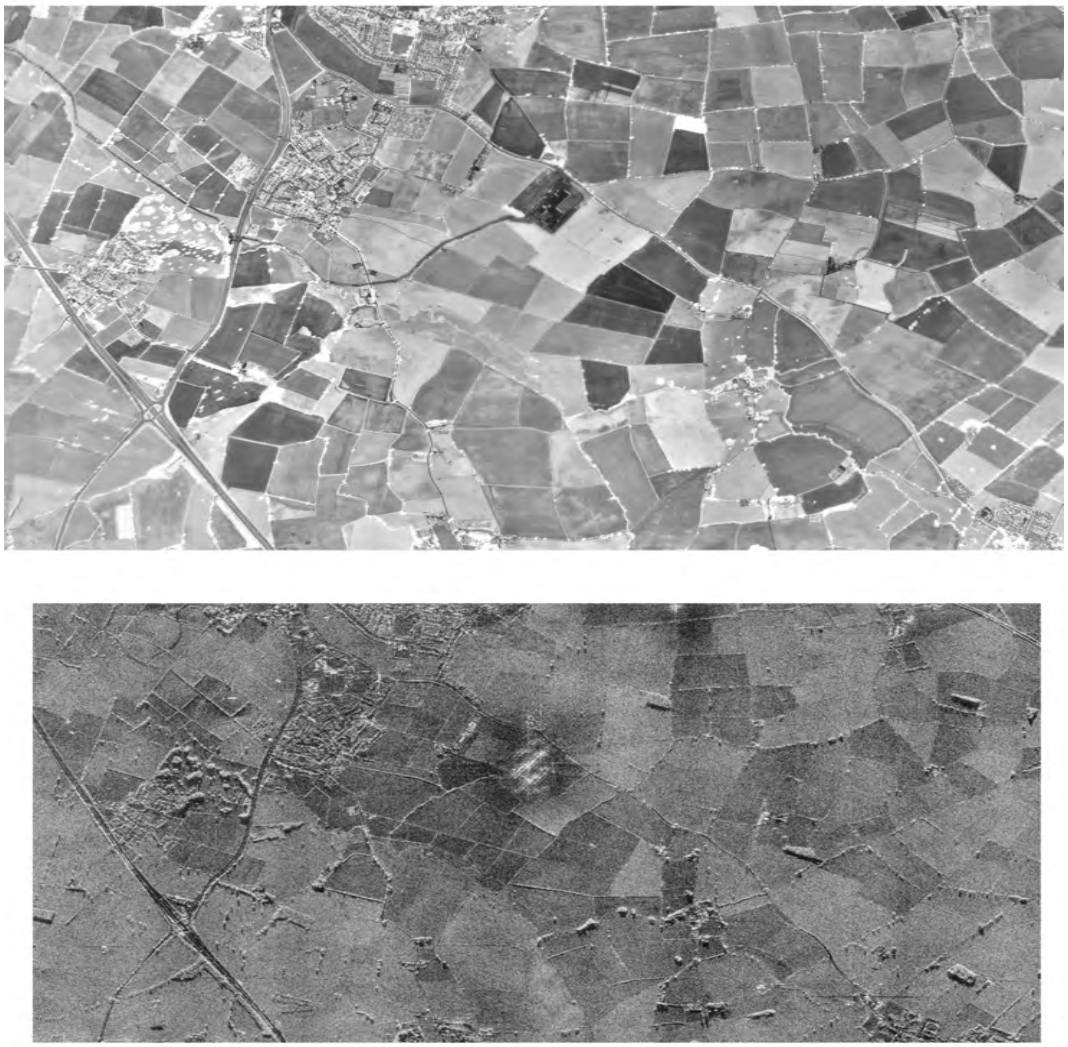

FIGURE 7 An incoherent optical image (above) and a coherent (Synthetic Aperture Radar) image of the same region of Northamptonshire, England.

or echo is recorded at fixed time intervals along the flight path.

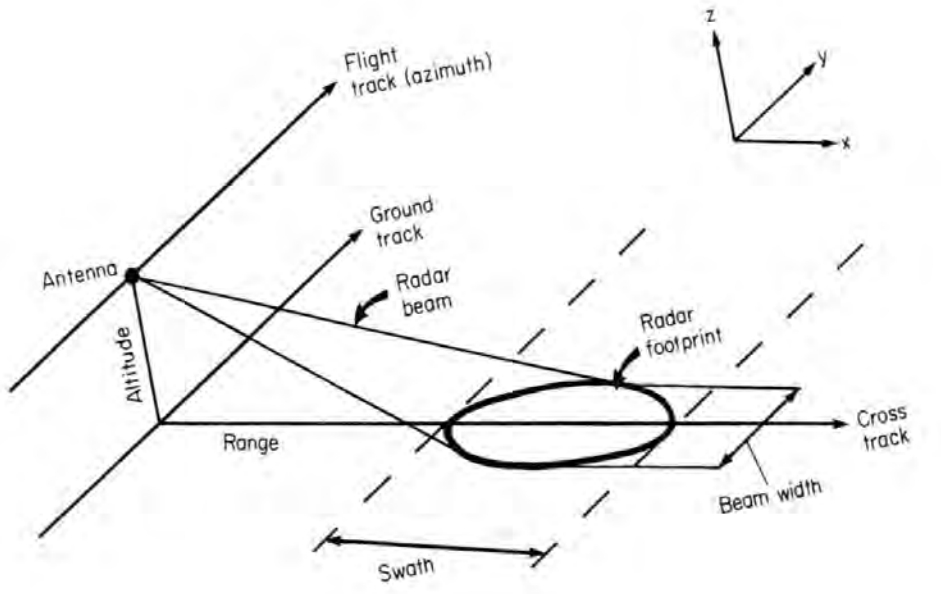

FIGURE 8 Basic geometry of an airborne SAR imaging system.

5.1.1 The Radar Pulse

SAR is a peak power limited system. In other words it operates at the maximum power available. The energy of the system is therefore given by

$$\text{Energy} = \text{Peakpower} \times \text{Time}$$

In order to transmit a microwave field with enough energy to establish a measurable return, the duration of the pulse must be made relatively long. The length of this pulse is large compared to the wavelength and, hence the system is based on application of a side-band spectrum. If a simple on/off pulse is emitted then the characteristic spectrum is a narrow-band sinc function. The frequency content of this type of pulse is not usually broad enough to obtain adequate range resolution. For this reason, a frequency sweep or 'chirp' is applied over the duration of the pulse. Even with a frequency sweep applied to it, the pulse has a very narrow frequency band. In other words, the energy of the pulse is concentrated near to the carrier frequency. The type of pulse that is actually used is given by (complex form)

$$p(\tau) = \exp(ik_0\tau)\exp(i\alpha\tau^2), \quad -T/2 \le \tau \le T/2$$

where T is the pulse length, τ is time $\times$ speed of light, α is the quadratic chirp rate / (speed of light)^2 and k_0 is the carrier wave number (carrier frequency $= \frac{k_0}{2\pi} \times$ speed of light). Note that in reality the pulse is of course not a complex but a real valued function of time. It is given by the real part of p, i.e. $\cos(k_0\tau + \alpha\tau^2)$. This type of pulse is just one of a number of different types of coded pulses that can in

principle be used. It is used extensively in radar systems because of its properties for generating range resolution and it can be implemented comparatively easily. The instantaneous phase of this pulse is $k_0\tau + \alpha\tau^2$. The rate of change of phase, or frequency modulation, is therefore $k_0 + 2\alpha\tau$ which is linear in τ. For this reason, the pulse is known as a linear frequency modulated (FM) chirp. In general, most SARs utilize values of k_0 and α where

$$k_0 >> 1$$

and

$$\alpha << 1.$$

For example, in the system used to produce the SAR image given in Figures 5.1,

$$k_0 \simeq 224\mathrm{m}^{-1}$$

and the quadratic chirp rate was $2\pi \times 10^{13}\mathrm{sec}^{-2}$ giving

$$\alpha \simeq 7 \times 10^{-4}\mathrm{m}^{-2}.$$

5.1.2 The Range Spectrum

The spectrum of the FM chirp is obtained by evaluating the integral

$$P(k) = \int_{-T/2}^{T/2} \exp(ik_0\tau)\exp(-i\alpha\tau^2)\exp(-ik\tau)d\tau \tag{5.1}$$

This is given by

$$P(k) = \sqrt{\frac{\pi}{2\alpha}}\left[K\left(\frac{\alpha T + u}{\sqrt{2\pi\alpha}}\right) + K\left(\frac{\alpha T - u}{\sqrt{2\pi\alpha}}\right)\right]\exp(-iu^2/4\alpha)$$

where $u = k + k_0$ and

$$K(x) = \int_0^x \exp(i\pi x^2/2)dx = C(x) + iS(x)$$

with real and imaginary parts

$$C(x) = \int_0^x \cos\frac{\pi}{2}x^2 dx$$

and

$$S(x) = \int_0^x \sin\frac{\pi}{2}x^2 dx.$$

The integrals above are known as **Fresnel integrals**. Figure 9 is a sketch of the real valued pulse $\cos(k_0\tau + \alpha\tau^2)$ and its characteristic amplitude spectrum. Observe that the bandwidth of the pulse is determined by the value of αT. With microwave systems this is typically two to three orders of magnitude smaller than the carrier wavenumber k_0.

5.1.3 Range Processing

Consider a single point scatterer which reflects a replica of the transmitted pulse. At the receiver the return signal is coherently mixed down to base-band (i.e. frequency demodulated). In practice, the field that is actually measured is of course not a complex but a real valued signal. The imaginary part of this signal is obtained using a quadrature filter which demodulates the return signal using $\sin(k_0\tau)$ instead of $\cos(k_0\tau)$. This is equivalent to computing the Hilbert transform of the signal after demodulation to base-band. The complex or analytic signal that is obtained after demodulation is given by

$$\exp(i\alpha\tau^2), \quad -T/2 \leq \tau \leq T/2.$$

At this stage, the range resolution is determined by the pulse length T. By applying

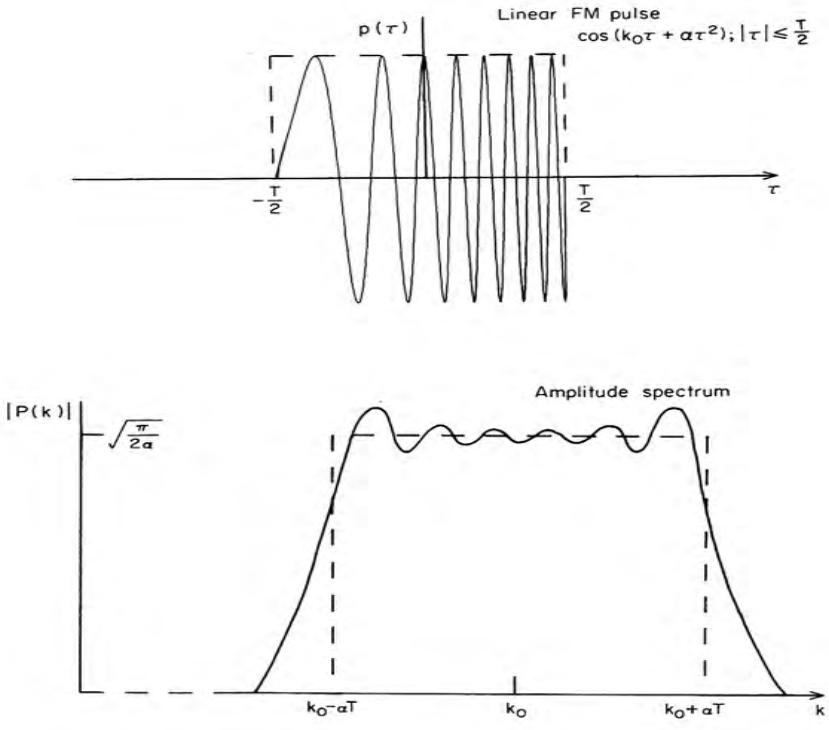

FIGURE 9 Sketch of a linear frequency modulated (chirped) pulse and its characteristic amplitude spectrum.

a suitable process to the return signal, we can enhance the range resolution and hence obtain a more accurate record of the position in range of the point scatterer. This is achieved by correlating the signal with its complex reference function $\exp(-i\alpha\tau^2)$. In SAR and other pulse-echo systems which utilize a linear FM pulse, this process is known as range compression. The range compressed data $R(\tau)$ can be written as (u being a dummy variable)

$$R(\tau) = \int_{-T/2}^{T/2} \exp[-i\alpha(\tau+u)^2]\exp(i\alpha u^2)du.$$

Expanding $(\tau + u)^2$, this equation becomes

$$R(\tau) = \exp(-i\alpha\tau^2) \int_{-T/2}^{T/2} \exp(-2i\alpha u\tau) du.$$

Evaluating the integral over u, we have

$$R(\tau) = T \exp(-i\alpha\tau^2) \text{sinc}(\alpha T\tau).$$

The length of the pulse T is relatively large. As a consequence of this, the sinc function is very narrow compared with the complex exponential. For this reason we have

$$\cos(\alpha\tau^2)\text{sinc}(\alpha T\tau) \simeq \text{sinc}(\alpha T\tau)$$

and

$$\sin(\alpha\tau^2)\text{sinc}(\alpha T\tau) \simeq 0.$$

The range compressed signal can therefore be written as

$$R(\tau) \simeq T\text{sinc}(\alpha T\tau), \quad T >> 1.$$

By defining the range resolution to be the distance between the first two zeros of the sinc function which occur when $\alpha T\tau = \pm\pi$ the range resolution is given by

$$\text{Range resolution} = 2\pi/\alpha T \text{ metres} \quad .$$

Observe that, as the value of αT increases, the range resolution improves. For a 20 μs pulse, $T = 6$ km and, with $\alpha = 7 \times 10^{-4}$ m^{-2}, the range resolution is approximately 1.5 metres.

5.1.4 Azimuth Processing

As the radar travels along its flight path (repeatedly emitting a linear FM pulse and recording the back-scattered electric field that is scattered by the ground), the radar beam illuminates an area of the ground which depends upon the grazing angle, its angle of divergence and the range at which the radar operates. The width of the beam in azimuth is given by $R\tan(\beta/2)$ where R is the range and β is the angle of divergence of the beam. For a SAR system, this value corresponds to the maximum length of the synthetic aperture as shown in Figure 10. In practice, $\beta \sim 1^0$ and so the width of the beam is approximately given by $R\beta/2$. This value determines the resolution in azimuth of the so called Real Aperture Radar or RAR. At a range of of say 50 km with $\beta = 1^0$, this resolution is just under a kilometre which is very poor and of little practical use. Hence real aperture radar images are only useful when short ranges are involved. The whole point of SAR is to obtain high resolution at long ranges. By studying the response of the radar in azimuth as it passes by a scatterer, we can synthesize the resolution via the principle demonstrated in Figure 11. If we consider the radar to be a point source then the field that is produced may therefore be described by the three-dimensional Green function. At relatively

large distances from the location of the source, the Green function can be simplified using the Fresnel approximation (see Appednix 1). This provides a description for the wavefield in the intermediate or Fresnel zone. The wavefronts in this zone have a curvature which is parabolic as illustrated in Figure 11. Using the geometry shown

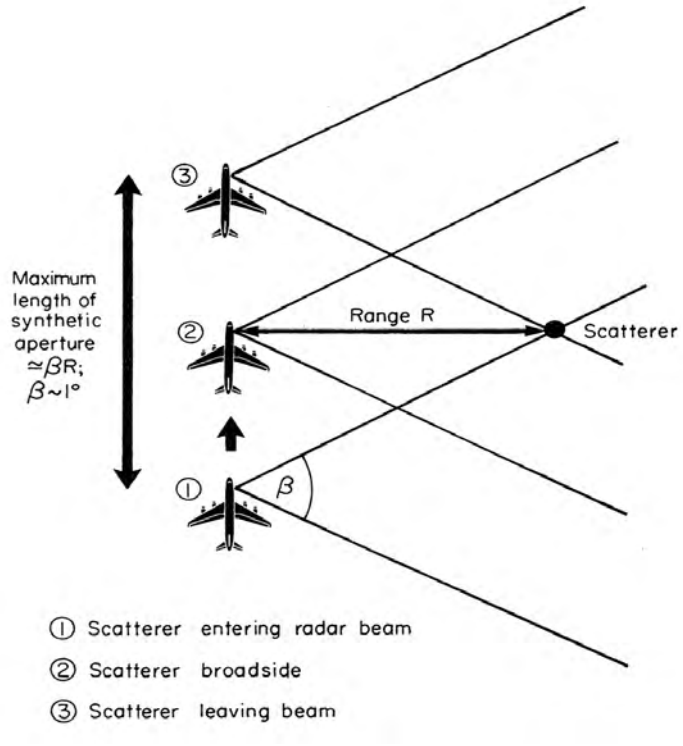

FIGURE 10 Plan view of a SAR showing the maximum length of the synthetic aperture.

in this figure, from Pythagoras' theorem we have

$$(R + \delta R)^2 + y^2 = R^2$$

or

$$2R\delta R + (\delta R)^2 + y^2 = 0.$$

If the angle of divergence of the beam is small, then δR is much less than 1. We can then ignore the nonlinear term $(\delta R)^2$ leaving the equation

$$2\delta R = -\frac{y^2}{R}. \tag{5.2}$$

A simple plane wave travelling along the two-way path length $2(R + \delta R)$ can therefore be written as

$$\exp[(-2ik_0(R + \delta R)] = \exp(-2ik_0 R)\exp(-2ik_0\delta R)$$

where k_0 is the wavenumber. This wave has two phase factors. The first phase $2k_0R$ is constant but the second phase $2k_0\delta R$ is, from equation (5.2) a function of y and is given by k_0y^2/R. Hence, as the radar moves past the scatterer a quadratic phase

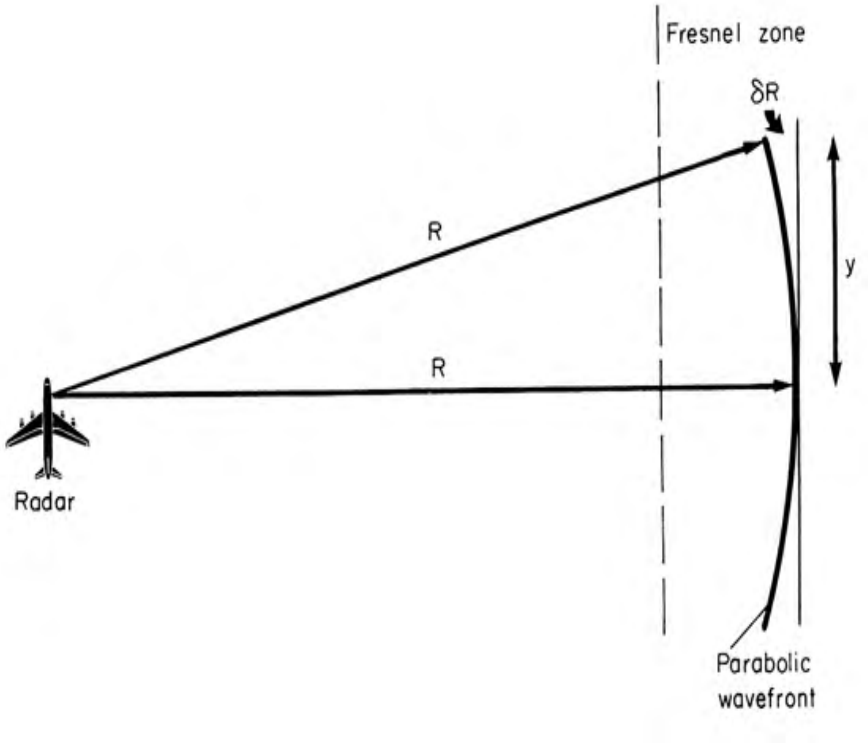

FIGURE 11 By the time the wavefield emitted by the radar has reached a point scatterer, the curvature of the wavefront is parabolic. Scattering occurs in the Fresnel zone. This gives a phase history that is proportional to the square of the distance moved in azimuth.

shift takes place. If we denote the width of the beam at R by L, then the complex azimuth response of the radar can be written as

$$\exp(ik_0y^2/R), \quad -L/2 \leq y \leq L/2$$

where $-L/2$ is the point where the scatterer enters the beam and $L/2$ is the position where the scatterer leaves the beam. A plot of the azimuth response of a SAR is given in Figure 12. In some cases, this response can be clearly observed with real data when the radar passes by a strong scatterer with a large radar cross-section. An example of this is given in Figure 13. If the beamwidth is small, then this effect is not significant. Also, only if k_0 is sufficiently large will the effect be observed. In other words, the wavelength of the wavefield must be small compared with the range.

The analysis above demonstrates that the azimuth response of the radar is the same as the response in range to a linear FM pulse. Hence, by utilizing the principles of range compression, we can enhance the azimuth resolution. This is known as azimuth compression and, like range compression, is based on correlating the complex function $\exp(ik_0y^2/R)$ with its complex reference function $\exp(-ik_0y^2/R)$ over the beam width L. Hence, the azimuth compressed signal is given by

$$A(y) = \int_{-L/2}^{L/2} \exp[-ik_0(y+u)^2/R] \exp(ik_0u^2/R)du.$$

Expanding $(y+u)^2$ and evaluating the integral over u, we get

$$\begin{aligned} A(y) &= L\exp(-ik_0y^2/R)\text{sinc}(k_0Ly/R) \\ &\simeq L\text{sinc}(k_0Ly/R), \quad L >> 1. \end{aligned}$$

For both azimuth compression and range compression, the correlation between the return signal and its reference may be computed in Fourier space using the correlation theorem and a FFT.

By defining the azimuth resolution to be the distance between the first zeros of the sinc function which occur when $k_0Ly/R = \pm\pi$, the azimuth resolution is given by

$$\text{Azimuth resolution} = 2\pi R/k_0L = 2\pi/\beta k_0 \text{ metres} \quad .$$

The microwave antenna (i.e. essentially the horn at the end of the microwave transmission line) acts like a rectangular aperture which diffracts an otherwise collimated beam of microwaves. The Kirchhoff diffraction integral for fixed $k = k_0$ is given by

$$\int_S \exp(i\mathbf{k}\cdot\mathbf{r}) \exp(-ik_0\hat{\mathbf{r}}_0\cdot\mathbf{r})d^2\mathbf{r}, \quad \hat{\mathbf{r}}_0 = \frac{\mathbf{r}_0}{|\,\mathbf{r}_0\,|}$$

which, for an aperture of width w, say, and an incident plane wave propagating in the z-direction (where $\mathbf{k} = \hat{\mathbf{z}}k_0$), becomes (with $r_0 \sim z_0$ and ignoring scaling)

$$\int_{-w/2}^{w/2}\int_{-w/2}^{w/2} \exp(-ik_0x_0x/z_0) \exp(-ik_0y_0y/z_0)dxdy$$

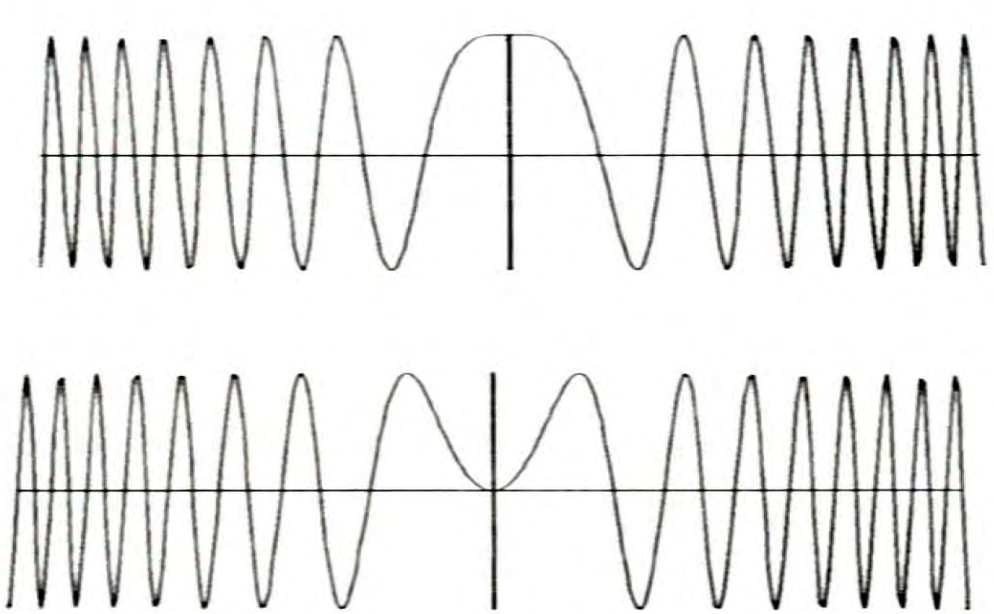

FIGURE 12 Real (top) and imaginary (bottom) components of the theoretical response in azimuth of a SAR to a single point scatterer, i.e. $\cos(k_0 y^2/R)$ and $\sin(k_0 y^2/R)$, respectively.

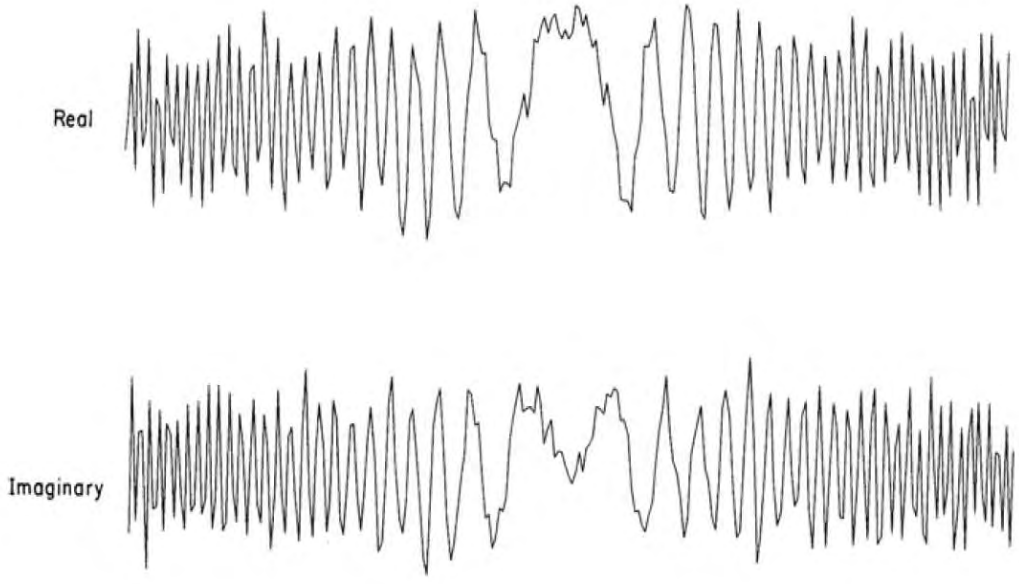

FIGURE 13 Example of the experimental response in azimuth of a SAR to a single point scatterer. It is clearly a noisy version of Figure 12.

$$= 4\frac{\sin(k_0 x_0 w/2z_0)}{k_0 x_0/z_0}\frac{\sin(k_0 y_0 w/2z_0)}{k_0 y_0/z_0}.$$

The first zeros of this diffraction pattern in azimuth can be taken to determine the width of the radar beam (i.e. the first lobe). These zeros occur when

$$k_0\frac{w}{2}\sin\frac{\beta}{2} = \pm\pi$$

where $\beta/2 = y_0/z_0$. Hence, for small values of β,

$$\beta \simeq \frac{4\pi}{k_0 w}$$

and, hence the azimuth resolution is proportional to w. The azimuth or synthetic resolution of the SAR is therefore independent of the wavelength.

By studying the response of the radar to a point scatterer in range, and then in azimuth, we have established the form of the SAR point spread function. This is given by

$$P(x,y) = LT\text{sinc}(\alpha T x)\text{sinc}(\beta k_0 y).$$

It is identical to the diffraction pattern produced by a rectangular aperture. Thus, the (post-processed) SAR image data $D(x,y)$ generated by scattering from the ground is given by the convolution of the object function for the ground $O(x,y)$ with the appropriate point spread function, i.e.

$$D(x,y) = P(x,y) \otimes\otimes O(x,y). \tag{5.3}$$

A SAR image is a grey level display of the amplitude modulations in the data, i.e.

$$I_{SAR}(x,y) = \mid D(x,y) \mid .$$

The object function describes the imaged properties of the ground surface. The conventional model for this function is the point scattering model. This is where the object function is taken to be a distribution of point scatterers each of which reflects a replica of the emitted pulse and can be written in the form

$$O(x,y) = \sum_i\sum_j \delta(x - x_i)\delta(y - y_j).$$

Here, nothing is said about the true physical nature of the ground surface such as its shape and material (dielectric) properties. In the following Section this shortcoming is addressed.

5.2 Scattering Model

By considering the response of the radar to a single point scatterer, the basic processing technique required to recover a SAR image can be established. However, this approach conveys no information about the possible physical interpretation of a SAR image. To do this the relationship between the object function and the physical properties of the ground surface such as its dielectric properties and height fluctuations must be established. In this Section approximate expressions for the object functions associated with different polarizations are derived.

5.2.1 A Physical Model for SAR

Consider the model illustrated in Figure 14. Here, x is the range coordinate, y is the azimuth coordinate and z is the vertical co-ordinate. Let the ground be composed of three-dimensional variations in the permittivity ϵ and conductivity σ with height variations h. We shall assume that the relative permeability of the ground is 1. Hence, the back-scattered field detected by the radar is produced by variations in $\epsilon(\mathbf{r})$ and $\sigma(\mathbf{r})$ over a region of space $\mathbf{r} = \hat{\mathbf{x}}x + \hat{\mathbf{y}}y + \hat{\mathbf{z}}z$ where $0 \leq z \leq h(x,y)$ and $\hat{\mathbf{x}}x + \hat{\mathbf{y}}y \in A$ - the area of the ground illuminated by the radar beam (i.e. the radar footprint). For $z > h(x,y)$, ϵ and σ are equal to the permittivity and conductivity of the atmosphere. The permittivity of the atmosphere is taken to be the same as for a vacuum and the conductivity of the atmosphere is assumed to be zero. Thus, if a denotes the altitude at which the radar operates, then for all values of z between h and a, $\epsilon = \epsilon_0$ and $\sigma = 0$. The field that is measured in a SAR is the electric field and so we can work with equations for the electric field alone. From Maxwell's equations, we can write the basic wave equation for this field in the form (see Chapter 2)

$$(\nabla^2 + k^2)\widetilde{\mathbf{E}} = -k^2\gamma\widetilde{\mathbf{E}} + ikz_0\sigma\widetilde{\mathbf{E}} - \nabla(\widetilde{\mathbf{E}} \cdot \nabla \ln \epsilon)$$

where

$$\gamma = \frac{\epsilon - \epsilon_0}{\epsilon_0},$$

k is the wavenumber and z_0 is the impedance of free space ($\simeq 376.6$ ohms).

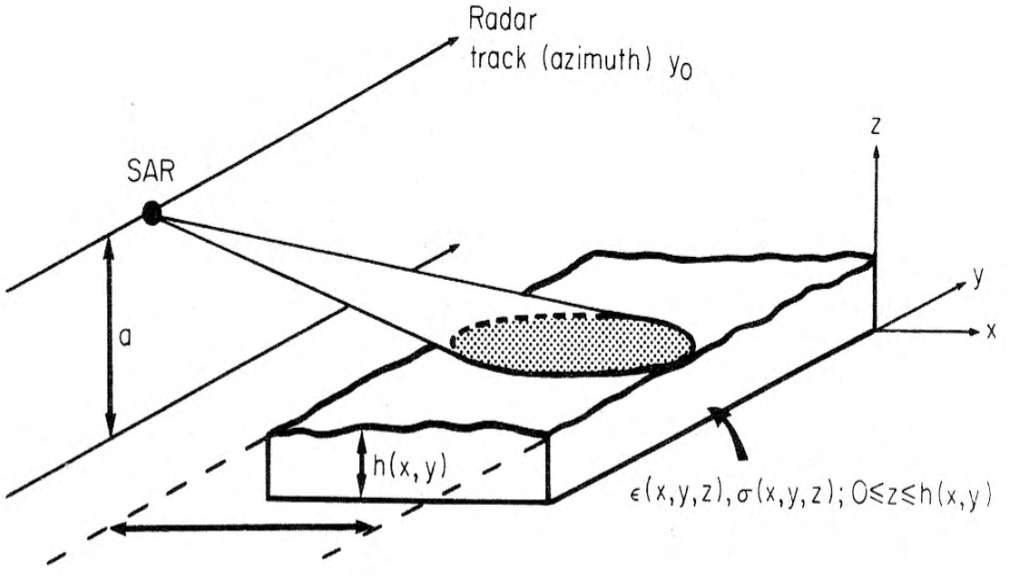

FIGURE 14 Physical model for an airborne SAR.

Assuming that the scattered field only weakly perturbs the incident field, i.e.

$$\|\widetilde{\mathbf{E}}_s\| << \|\widetilde{\mathbf{E}}_i\|$$

and writing $\epsilon = \epsilon_r \epsilon_0$ where ϵ_r is the relative permittivity, we obtain

$$(\nabla^2 + k^2)\widetilde{\mathbf{E}}_s = -k^2\gamma\widetilde{\mathbf{E}}_i + ikz_0\sigma\widetilde{\mathbf{E}}_i - \nabla(\widetilde{\mathbf{E}}_i \cdot \nabla \ln \epsilon_r) \tag{5.4}$$

where

$$\gamma = \epsilon_r - 1.$$

Note that, in this model, the effects of using different polarizations are determined entirely by the term $\nabla(\widetilde{\mathbf{E}}_i \cdot \nabla \ln \epsilon_r)$. If this term is neglected, then the behaviour of the electric field is independent of its polarization (i.e. the wave equation remains the same when the polarization of the electric field is changed). Our problem is to solve equation (5.4) for the scattered electric field $\widetilde{\mathbf{E}}_s$ and to write the solution in a form that is the same as equation (5.3) so that the object function can be defined in terms of the physical properties of the ground (ϵ_r, σ and h). To do this we need a suitable model for the incident field.

5.2.2 Green's Function for Airborne SAR

Consider the radar to be a point source. We may then consider a model for the incident field of the form

$$\widetilde{\mathbf{E}}_i = \hat{n} P g$$

where P is the spectrum of the pulse that the radar emits given by equation (5.1) and g is the three-dimensional 'out-going' Green's function given by

$$g(\mathbf{r} \mid \mathbf{r}_0, k) = \frac{\exp(ik \mid \mathbf{r} - \mathbf{r}_0 \mid)}{4\pi \mid \mathbf{r} - \mathbf{r}_0 \mid}.$$

The geometry of an airborne SAR allows us to approximate the Green's function. Writing the path length $\mid \mathbf{r} - \mathbf{r}_0 \mid$ in Cartesian coordinates,

$$\mid \mathbf{r} - \mathbf{r}_0 \mid = (x - x_0)\left(1 + \frac{(y - y_0)^2}{(x - x_0)^2} + \frac{(z - z_0)^2}{(x - x_0)^2}\right)^{1/2}$$

and employing the conditions

$$\frac{(y - y_0)^2}{(x - x_0)^2} << 1$$

and

$$\frac{(z - z_0)^2}{(x - x_0)^2} << 1$$

a binomial expansion gives

$$\mid \mathbf{r} - \mathbf{r}_0 \mid \simeq x - x_0 + \frac{1}{2}\frac{(y - y_0)^2}{(x - x_0)} + \frac{1}{2}\frac{(z - z_0)^2}{(x - x_0)}.$$

This result yields an expression for the Green's function in the Fresnel zone. In this case, we retain terms which are quadratic in both the azimuth and vertical directions. It is the inclusion of quadratic terms of this type which forms the theoretical basis

for synthetic aperture imaging. Physically, we are assuming that the wavefront as a function of y and z has a curvature which is parabolic. The conditions required to do this place limits on the grazing angle θ and the angle of divergence of the radar beam β. In terms of θ and β, we may write these conditions in the form

$$\tan^2(\beta/2) << 1$$

and

$$\tan^2\theta << 1.$$

It is reasonable to restrict values of θ and β to being less than or equal to 10^o when $\tan^2\theta$ and $\tan^2(\beta/2)$ are two and three orders of magnitude less than 1, respectively. This upper limit for θ and β falls well within the values of these parameters that are used in airborne SAR systems, where θ is typically $5^o - 10^o$ and $\beta \sim 1^o$. The above expression for the path length can be further simplified by exploiting the fact that the range x_0 at which the radar operates is large compared to the width of ground that is illuminated by the beam (the swath width), i.e. we can introduce the condition

$$\frac{\mid x \mid}{x_0} << 1.$$

This allows us to write

$$\mid \mathbf{r} - \mathbf{r}_0 \mid = x - x_0 - \frac{(y-y_0)^2}{2x_0} - \frac{(z-z_0)^2}{2x_0}.$$

The Green's function is then given by

$$g = \frac{1}{4\pi r_0}\exp[ik(x-x_0)]\exp[-ik(y-y_0)^2/2x_0]\exp[-ik(z-z_0)^2/2x_0].$$

The parameter r_0 remains fixed throughout the operation of the SAR and is known as the slant range (i.e. the distance between the radar and the scattering region).

5.2.3 Wave Equations for SAR

Let us consider a SAR that can emit a vertically polarized electric field of the form

$$\widetilde{\mathbf{E}}_i = (\hat{\mathbf{z}}\cos\theta + \hat{\mathbf{x}}\sin\theta)Pg \tag{5.5}$$

or a horizontally polarized electric field where $\mathbf{E}_i$ is given by

$$\widetilde{\mathbf{E}}_i = \hat{\mathbf{y}}Pg. \tag{5.6}$$

Substituting equation (5.6) into equation (5.4) and taking the dot product of each term with $\hat{\mathbf{y}}$, the behaviour of the HH scattered field U_{HH} is determined by the wave equation

$$\begin{aligned}(\nabla^2 + k^2)U_{HH} &= -k^2\gamma Pg + ikz_0\sigma Pg \\ &-\frac{\partial}{\partial y}\left(Pg\frac{\partial}{\partial y}\ln\epsilon_r\right), \quad U_{HH} = \hat{\mathbf{y}}\cdot\widetilde{\mathbf{E}}_s.\end{aligned} \tag{5.7}$$

The cross polarized scattered field in this case is obtained by taking the dot product of each term with $\hat{\mathbf{z}}\cos\theta + \hat{\mathbf{x}}\sin\theta$ giving

$$(\nabla^2 + k^2)U_{HV} = -\left(\cos\theta\frac{\partial}{\partial z} + \sin\theta\frac{\partial}{\partial x}\right)\left(Pg\frac{\partial}{\partial y}\ln\epsilon_r\right),$$

$$U_{HV} = (\hat{\mathbf{z}}cos\theta + \hat{\mathbf{x}}\sin\theta)\cdot\widetilde{\mathbf{E}}_s. \tag{5.8}$$

In a similar way, the wave equations for the VV and VH scattered fields are obtained by substituting equation (5.5) into equation (5.4) and taking the dot product of each term with $\hat{\mathbf{z}}\cos\theta + \hat{\mathbf{x}}\sin\theta$ and $\hat{\mathbf{y}}$, respectively. We then obtain

$$(\nabla^2 + k^2)U_{VV} = -k^2\gamma Pg + ikz_0\sigma Pg$$

$$-\left(\cos\theta\frac{\partial}{\partial z} + \sin\theta\frac{\partial}{\partial x}\right)\left(\cos\theta Pg\frac{\partial}{\partial z}\ln\epsilon_r + \sin\theta Pg\frac{\partial}{\partial x}\ln\epsilon_r\right),$$

$$U_{VV} = (\hat{\mathbf{z}}\cos\theta + \hat{\mathbf{x}}\sin\theta)\cdot\widetilde{\mathbf{E}}_s \tag{5.8}$$

and

$$(\nabla^2 + k^2)U_{VH} = -\frac{\partial}{\partial y}\left(\cos\theta Pg\frac{\partial}{\partial z}\ln\epsilon_r + \sin\theta Pg\frac{\partial}{\partial x}\ln\epsilon_r\right),$$

$$U_{VH} = \hat{\mathbf{y}}\cdot\widetilde{\mathbf{E}}_s. \tag{5.9}$$

Notice that the behaviour of the VV and HH fields is determined by variations in both the permittivity and conductivity whereas that of the HV and VH fields depends on variations in the permittivity alone. This result immediately suggests a method of quantitative imaging with SAR. By measuring U_{VH} we can in principle determine ϵ_r, and therefore $\gamma(=\epsilon_r - 1)$. Hence by measuring U_{VV}, with γ and ϵ_r known, we can determine σ.

In general, fluctuations in ϵ_r and therefore $\ln\epsilon_r$ (as a function x, y and z) occur on a scale that is much smaller than the wavelength. For this reason we can write

$$\frac{\partial}{\partial u}\left(g\frac{\partial}{\partial v}\ln\epsilon_r\right) \simeq g\frac{\partial^2}{\partial u\partial v}\ln\epsilon_r$$

where both u and v are equal to x, y or z. For example, from equation (5.7)

$$\frac{\partial}{\partial y}\left(g\frac{\partial}{\partial y}\ln\epsilon_r\right) = g\frac{\partial^2}{\partial y^2}\ln\epsilon_r + \frac{\partial g}{\partial y}\frac{\partial}{\partial y}\ln\epsilon_r$$

$$= g\left(\frac{\partial^2}{\partial y^2}\ln\epsilon_r - ik_0\frac{(y-y_0)}{x_0}\frac{\partial}{\partial y}\ln\epsilon_r\right)$$

$$\simeq g\frac{\partial^2}{\partial y^2}\ln\epsilon_r$$

provided

$$L_y << \frac{x_0}{k_0 \mid y - y_0 \mid}$$

where L_y is the characteristic scale length over which variations in $\ln \epsilon_r$ occur. For an X-band radar operating at a range of 50 km with a beamwidth of 1 km,

$$L_y << 22\text{cm}$$

which is physically reasonable. This result allows us to reduce equation (5.7) and write it in the form

$$(\nabla^2 + k^2)U_{HH} \simeq -k^2\gamma Pg + ikz_0\sigma Pg - Pg\frac{\partial^2}{\partial y^2}\ln\epsilon_r. \tag{5.10}$$

Similarly, equations (5.8)-(5.9) become

$$(\nabla^2 + k^2)U_{HV} \simeq -\cos\theta Pg\frac{\partial^2}{\partial z\partial y}\ln\epsilon_r - \sin\theta Pg\frac{\partial^2}{\partial x\partial y}\ln\epsilon_r \tag{5.11}$$

$$\begin{aligned}(\nabla^2 + k^2)U_{VV} \simeq & -k^2\gamma Pg + ikz_0\sigma Pg - \cos^2\theta Pg\frac{\partial^2}{\partial z^2}\ln\epsilon_r \\ & -2\cos\theta\sin\theta Pg\frac{\partial^2}{\partial z\partial x}\ln\epsilon_r - \sin^2\theta Pg\frac{\partial^2}{\partial x^2}\ln\epsilon_r\end{aligned} \tag{5.12}$$

$$(\nabla^2 + k^2)U_{VH} \simeq -\cos\theta Pg\frac{\partial^2}{\partial y\partial z}\ln\epsilon_r - \sin\theta Pg\frac{\partial^2}{\partial y\partial x}\ln\epsilon_r. \tag{5.13}$$

5.2.4 Determination of the Back-scattered Fields

Now that a set of wave equations has been derived, we can concentrate on developing a solution for the back-scattered field that is observed by the radar. To start with, we shall develop a solution for the HH field. For the time being, let us consider the reduced wave equation

$$(\nabla^2 + k^2)U = -k^2 Pg\gamma + ikz_0 Pg\sigma \tag{5.14}$$

After demonstrating the basic analytical method we shall return to equations (5.10)-(5.13). Remember, we are aiming at a solution for the processed SAR data which gives a mathematical expression for the object function in terms of ϵ_r, σ and h.

The Green's function solution to equation (5.14) for the back-scattered field is

$$U = P\int (k^2\gamma - ikz_0\sigma)g^2 d^3\mathbf{r}. \tag{5.15}$$

The radar measures the back-scattered field at a fixed range x_0 and altitude z_0 over a finite distance in azimuth. Denoting the fixed range and altitude by R and a, respectively, the kernel of equation (5.15) becomes

$$g^2 = \frac{1}{16\pi^2 r_0^2}\exp[2ik(x-R)]\exp[-ik(y-y_0)^2/R]\exp[-ik(z-a)^2/R].$$

Writing

$$X = x - R, \quad Y = y - y_0 \quad \text{and} \quad Z = z - a$$

the back-scattered field as a function of y_0 and k is given by

$$U(y_0,k) = \frac{P}{16\pi^2 r_0^2}\int\int\int \exp[ik(2X - Y^2/R - Z^2/R)](k^2\gamma - ikz_0\sigma)dxdydz.$$

Because the bandwidth of the pulse is so small compared to the carrier frequency we can write $k^2\gamma$ and $ikz_0\sigma$ as $k_0^2\gamma$ and $ik_0z_0\sigma$, respectively. By taking the inverse Fourier transform of the integral equation above, the back-scattered field can be written in terms of its measured time history $u(y_0,\tau)$ at different points in azimuth y_0. Using the convolution theorem we then obtain

$$u(y_0,\tau) = \frac{1}{16\pi^2 r_0^2}\int\int\int p(\tau + 2X - Y^2/R - Z^2/R)(k_0^2\gamma - ik_0z_0\sigma)dxdydz$$

where

$$u(y_0,\tau) = \frac{1}{2\pi}\int_{-\infty}^{\infty} U(y_0,k)\exp(ik\tau)dk.$$

The pulse is of the form

$$p(\tau) = \exp(ik_0\tau)\exp(i\alpha\tau^2).$$

Noting that $k_0 >> 1$ and $\alpha << 1$, by comparing the magnitude of terms which make up the kernel p we obtain

$$p(\tau + 2X + Z^2/R - Y^2/R) \simeq p(\tau + 2X)\exp(-ik_0Y^2/R)\exp(-ik_0Z^2/R).$$

This simplification is a consequence of the result

$$k_0 - \alpha(Y^2 + Z^2)/R \simeq k_0$$

and allows the scattered field to be written as

$$U(y_0,\tau) = \int\int p(\tau + 2X)\exp(-ik_0Y^2/R)f(x,y)dxdy$$

where f is the scattering function given by

$$f(x,y) = \frac{1}{16\pi^2 r_0^2}\int_0^h (k_0^2\gamma - ik_0z_0\sigma)\exp[-ik_0(z-a)^2/R]dz.$$

We now have an integral equation where our processing variables τ and y_0 have been separated into two different functions. This is why SAR data can be processed in range and azimuth separately. A further simplification can now be made to f by noting that

$$\frac{z}{a} << 1, \quad 0 \le z \le h$$

for an airborne SAR so that

$$(z-a)^2 = z^2 - 2za + a^2 \simeq -2za + a^2.$$

Hence, since $a/R = \tan\theta$, where θ is the grazing angle, the scattering function can be written as

$$f(x,y) = \frac{1}{16\pi^2 r_0^2}\exp(-ik_0 a\tan\theta)\int_0^h (k_0^2\gamma - ik_0 z_0\sigma)\exp(2ik_0 z\tan\theta)dz.$$

We now introduce a couple of tricks which are designed entirely to write the scattered field in a more convenient form. First of all we use the properties of the delta function to write

$$\int\int p(\tau+2X)\exp(-ik_0Y^2/R)f(x,y)dxdy$$

$$= \int_{-\infty}^{\infty} d\tau' p(\tau')\int\int \delta(\tau'-\tau-2X)\exp(-ik_0Y^2/R)f(x,y)dxdy$$

$$= \int d\tau' p(\tau')\int f[\tau'/2-\tau/2+R,y]\exp(-ik_0Y^2/R)dy.$$

Next, we let $x = 2R-\tau$ and $x' = \tau'+x$. Then, $\tau' = x'-x$, $\tau'/2-\tau/2+R = x'/2$ and $d\tau' = dx'$ and the scattered field can be written in the form

$$u(y_0,x) = \int dx' p(x'-x)\int f(x',y)\exp[-ik_0(y-y_0)^2/R]dy.$$

To be consistent with the notation now being used for the range variable we write y as y' and y_0 as y. The scattered field can then be written as

$$u(x,y) = \int\int \exp[ik_0(x'-x)]\exp[i\alpha(x'-x)^2]\exp[-ik_0(y'-y)^2/R]$$

$$\times f(x',y')dx'dy'.$$

At the receiver the scattered field, modelled by the above equation, is coherently mixed down to base-band. This is equivalent to multiplying it by $\exp(ik_0x)$ and provides the data

$$d(x,y) = \exp(ik_0x)u(x,y)$$

$$= \int\int \exp[i\alpha(x'-x)^2]\exp[-ik_0(y'-y)^2/R]\exp(ik_0x')f(x',y')dx'dy'.$$

This is a 2D convolution integral, and so we may write

$$d(x,y) = \exp(i\alpha x^2)\exp(-ik_0y^2/R)\otimes\otimes O(x,y)$$

where O is the object function given by

$$O(x,y) = \exp(ik_0x)f(x,y).$$

We can now apply the processing method which was explained in Section 10.2. Correlating these data with the functions $\exp(-i\alpha x^2)$ and $\exp(ik_0y^2/R)$ over the pulse length T and beam width L, respectively, we obtain

$$D(x,y) = \beta RT\,\mathrm{sinc}(\alpha Tx)\,\mathrm{sinc}(\beta k_0 y)\otimes\otimes O(x,y) \tag{5.16}$$

where

$$D(x,y) = d(x,y) \odot \odot \exp(-i\alpha x^2) \exp(ik_0 y^2/R)$$

and $\beta(= L/R)$ is the angle of divergence of the beam. The SAR image in then given by

$$I_{SAR}(x,y) = \mid D(x,y) \mid = \beta RT \mid \mathrm{sinc}(\alpha Tx)\mathrm{sinc}(\beta k_0 y) \otimes \otimes O(x,y) \mid .$$

Observe that this equation for D is the same as equation (5.3). However, in this case, the object function is defined in terms of a scattering function for the ground f.

By taking the two-dimensional Fourier transform of $D(x,y)$, equation (5.16) can then be written in $k_x k_y$-space as

$$\widetilde{D}(k_x,k_y) = \frac{\pi^2 R}{\alpha k_0} F(k_x - k_0, k_y); \quad -\alpha T \leq k_x \leq \alpha T, \quad -\beta k_0 \leq k_y \leq \beta k_0 \qquad (5.17)$$

where

$$\widetilde{D}(k_x,k_y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} D(x,y) \exp(-ik_x x) \exp(-ik_y y) dx dy$$

and

$$F(k_x - k_0, k_y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \exp(ik_0 x) f(x,y) \exp(-ik_x x) \exp(-ik_y y) dx dy.$$

From equation (5.17), it is clear that range compression provides a sample of the spectrum F of width $2\alpha T$ located at k_0. Unlike the range spectrum the azimuth spectrum is not the result of a spectral shift from GHz to MHz. The azimuth spectrum therefore gives base-band information on the nature of the scattering function f band limited by $2\beta k_0$. The spectral content of f that is acquired is therefore a rectangle of area $4\alpha\beta k_0 T$ centred on $(-k_0, 0)$ in $k_x k_y$ space. This is shown in Figure 15 which illustrates that the spectral information (in contrast to resolution) on the ground depends on the wavelength of the microwaves. The wavelength determines the characteristic scale length over which scattering takes place. This leads to a marked difference between SAR images obtained at different wavelengths. An example of this is shown in Figure 16 which compares an XVV and LVV SAR image of the same region.

Let us now return to equations (5.10)-(5.13). Recall that we worked with the reduced wave equation (5.14) in order to demonstrate the basic analytic method. Now that this has been done we are in a position to go back and repeat the calculation for equations (5.10)-(5.13). From equation (5.10), the back-scattered HH field is

$$U_{HH} = P \int \int \int \left(k^2 \gamma - ikz_0 \sigma + \frac{\partial^2}{\partial y^2} \ln \epsilon_r \right) g^2 dx dy dz.$$

This equation is identical in form to equation (5.15). The processed SAR data can therefore be written without further proof as

$$D_{HH}(x,y) = TR\beta \mathrm{sinc}(\alpha Tx)\mathrm{sinc}(\beta k_0 y) \otimes \otimes O_{HH}(x,y)$$

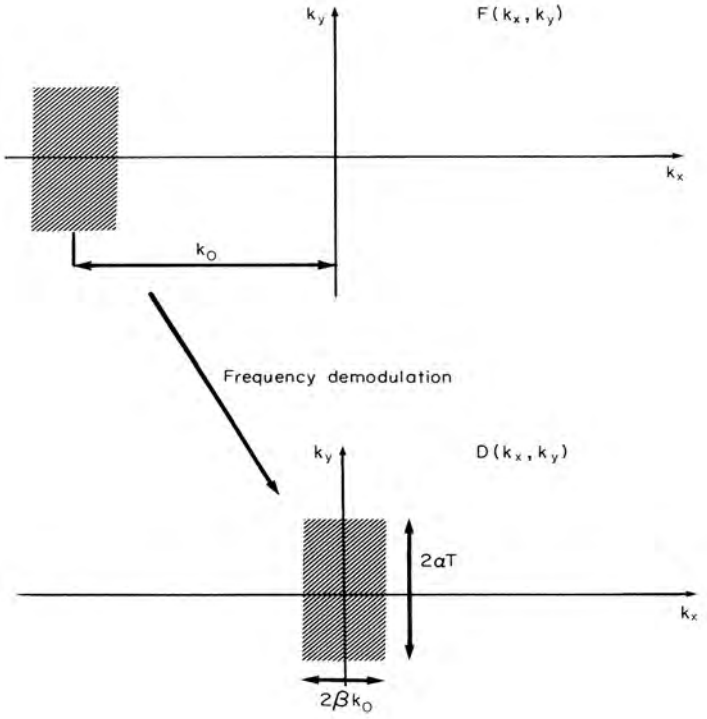

FIGURE 15 The shaded region represents the band of the spatial frequencies on the scattering function for the ground truth that is obtained with a SAR.

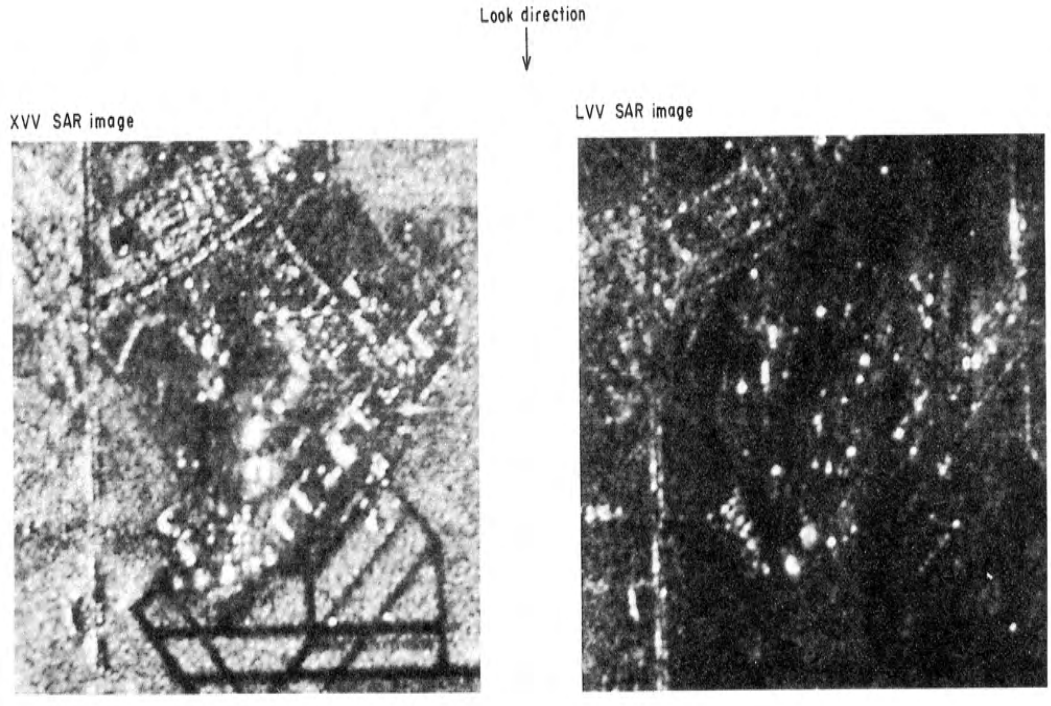

FIGURE 16 Comparison of two SAR images of the same region using different wavlengths: $\lambda = 2.8$ cm (left) and $\lambda = 24$cm (right).

where the HH object function is given by

$$O_{HH} = \frac{1}{16\pi^2 r_0^2} \exp(-ik_0 a \tan\theta) \exp(ik_0 x)$$

$$\times \int_0^h \left(k_0^2 \gamma - ik_0 z_0 \sigma + \frac{\partial^2}{\partial y^2} \ln \epsilon_r \right) \exp(2ik_0 z \tan\theta) dz. \tag{5.18}$$

A similar type of model can be generated for different polarization data. To begin with, we can evaluate the cross polarized scattered field. From equation (5.11), the back-scattered HV field is given by

$$U_{HV} = P \int\int\int \left(\cos\theta \frac{\partial^2}{\partial z \partial y} \ln \epsilon_r + \sin\theta \frac{\partial^2}{\partial x \partial y} \ln \epsilon_r \right) g^2 dx dy dz.$$

Once again the form of this equation is identical to that of equation (5.15). Hence the processed HV SAR data is

$$D_{HV}(x,y) = TR\beta \text{sinc}(\alpha T x) \text{sinc}(\beta k_0 y) \otimes\otimes O_{HV}(x,y)$$

where the HV object function is given by

$$O_{HV} = \frac{1}{16\pi^2 r_0^2} \exp(-ik_0 a \tan\theta) \exp(ik_0 x)$$

$$\times \int_0^h \left(\cos\theta \frac{\partial^2}{\partial z \partial y} \ln \epsilon_r + \sin\theta \frac{\partial^2}{\partial x \partial y} \ln \epsilon_r \right) \exp(2ik_0 z \tan\theta) dz. \tag{5.19}$$

From equation (5.12) it is easy to show that D_{VV} is given by

$$D_{VV}(x,y) = TR\beta \text{sinc}(\alpha T x) \text{sinc}(\beta k_0 y) \otimes\otimes O_{VV}(x,y)$$

where

$$O_{VV} = \frac{1}{16\pi^2 r_0^2} \exp(-ik_0 a \tan\theta) \exp(ik_0 x)$$

$$\times \int_0^h \left(k_0^2 \gamma - ik_0 z_0 \sigma + 2\cos\theta \sin\theta \frac{\partial^2}{\partial z \partial x} \ln \epsilon_r \right.$$

$$\left. + \sin^2\theta \frac{\partial^2}{\partial x^2} \ln \epsilon_r + \cos^2\theta \frac{\partial^2}{\partial z^2} \ln \epsilon_r \right) \exp(2ik_0 z \tan\theta) dz. \tag{5.20}$$

Finally, from equation (5.13), we get

$$D_{VH}(x,y) = TR\beta \text{sinc}(\alpha T x) \text{sinc}(\beta k_0 y) \otimes\otimes O_{VH}(x,y)$$

where

$$O_{VH} = \frac{1}{16\pi^2 r_0^2} \exp(-ik_0 z \tan\theta) \exp(ik_0 x)$$

$$\times \int_0^h \left(\cos\theta \frac{\partial^2}{\partial y \partial z} \ln \epsilon_r + \sin\theta \frac{\partial^2}{\partial y \partial x} \ln \epsilon_r \right) \exp(i2ik_0 z \tan\theta) dz. \tag{5.21}$$

5.3 The 'Sea Spikes' Problem

SAR images are highly sensitive to the polarization of the field that is emitted or received. In principle, this result can be used to classify regions of an image when it is known, *a priori*, how certain types of terrain affect different polarized radiation. One of the most dramatic effects occurs when microwaves are scattered by the sea surface at low grazing incidence. An example of this is shown in Figure 17. This figure

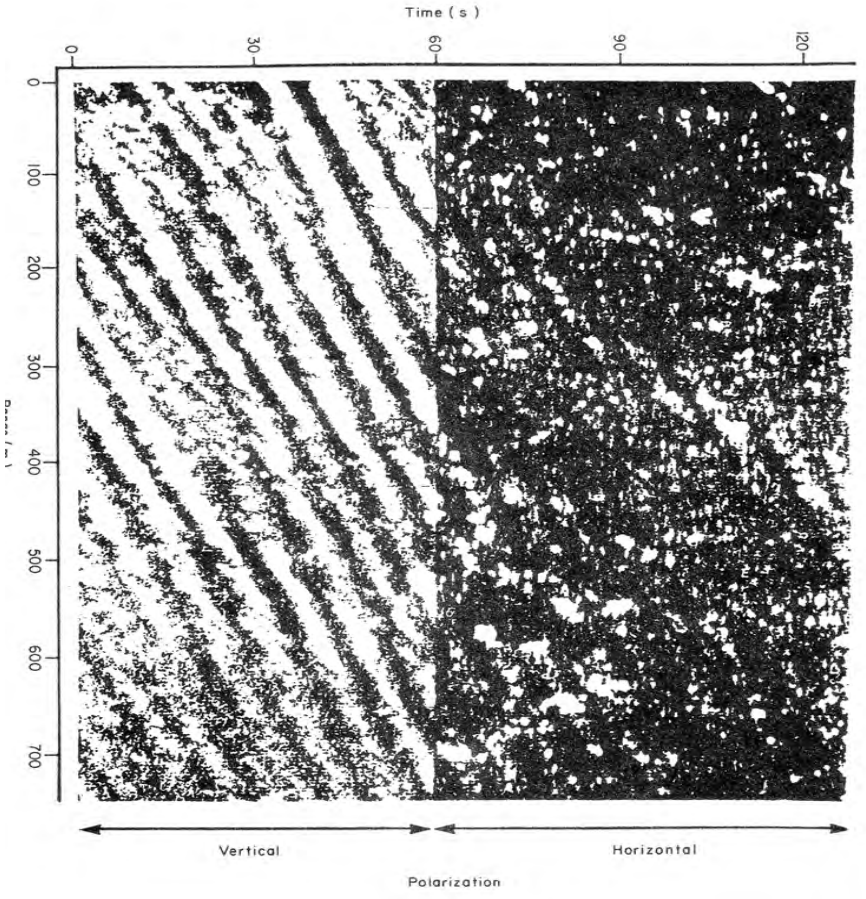

FIGURE 17 Real aperture radar images of the sea surface using vertical (left) and horizontal (right polarization).

shows two real aperture radar or RAR images of the sea surface using X-band HH and VV polarization. In this example, a pulse is emitted in a fixed time interval and the VV return measured over a set period of time (approximately 60 seconds). The radar is then switched to HH mode. Clearly, there is a marked difference between the two images. The VV image shows features which are due to reflections from the crests of waves that are aligned along the direction of the prevailing wind. These features are almost completely lost in the HH image, although it is just possible to observe the direction of wave motion. The HH image is dominated by a number of very intense reflections which are known as 'sea spikes'. This is a good example of a problem in image understanding. To explain this effect and solve the 'sea

spikes' problem we need to establish the physics associated with polarization and establish a suitable model for the sea surface. From previous results, under the Born approximation, polarization effects are characterized by the term $\nabla(\ln\epsilon_r \cdot \mathbf{E}_i)$ in the wave equation for the electric field. Hence, a good starting point is to investigate the characteristics of this term subject to a simplified model of the sea surface.

To a good approximation, the sea is a homogeneous conductive dielectric of varying height $h(x,y)$. We may therefore consider a model where

$$\epsilon_r(x,y,z) = \epsilon_{ro}, \quad z \leq h(x,y);$$

$$\sigma(x,y,z) = \sigma_0, \quad z \leq h(x,y)$$

and where

$$\left[\frac{\partial \epsilon_r}{\partial x}\right]_{z<h} = \left[\frac{\partial \epsilon_r}{\partial y}\right]_{z<h} = \left[\frac{\partial \epsilon_r}{\partial z}\right]_{z<h} = 0.$$

Typical values for ϵ_{ro} and σ_0 are 81 and 4.3 siemens/metre, respectively. In this case, for an X-band radar ($k_0 \simeq 224m^{-1}$), $k_0^2\gamma_0 \simeq 4 \times 10^6 m^{-2}$ and $k_0 z_0 \sigma_0 \simeq 3.6 \times 10^5$ so that

$$k_0^2\gamma_0 - ik_0 z_0 \sigma_0 \simeq k_0^2\gamma_0.$$

A simple mathematical model for the VV and HH RAR images given in Figure 17 can be obtained by letting the grazing angle θ approach zero. All terms involving $\sin\theta$ can then be neglected, giving

$$I_{RAR}^{ij}(x,y) = T \mid \mathrm{sinc}(\alpha T x) \exp(-ik_0 y^2/R) \otimes \otimes O_{ij}(x,y) \mid$$

where from equations (5.20) and (5.18),

$$O_{VV} = \frac{1}{16\pi^2 R^2} \exp(ik_0 x) \int_0^h \left(k_0^2\gamma_0 + \frac{\partial^2}{\partial z^2} \ln\epsilon_r\right) dz, \quad \gamma_0 = \epsilon_{ro} - 1$$

and

$$O_{HH} = \frac{1}{16\pi^2 R^2} \exp(ik_0 x) \int_0^h \left(k_0^2\gamma_0 + \frac{\partial^2}{\partial y^2} \ln\epsilon_r\right) dz$$

respectively. The VV object function is easy to evaluate, giving

$$O_{VV} = \frac{1}{16\pi^2 R^2} \exp(ik_0 x) \left(k_0^2\gamma_0 h + \frac{1}{\epsilon_{ro}} \left[\frac{\partial \epsilon_r}{\partial z}\right]_{z=h}\right).$$

The HH object function can be evaluated by using Leibniz' formula for the integral of a derivative, i.e.

$$\int_{a(x)}^{b(x)} \frac{\partial}{\partial x} f(x,y)dy = \frac{\partial}{\partial x} \int_{a(x)}^{b(x)} f(x,y)dy$$

$$+ \Big[f(x,y)\Big]_{y=a(x)} \frac{da}{dx} - \Big[f(x,y)\Big]_{y=b(x)} \frac{db}{dx}.$$

We then obtain

$$O_{HH} = \frac{1}{16\pi^2 R^2} \exp(ik_0 x) \left(k_0^2 \gamma_0 h - \frac{1}{\epsilon_{ro}} \left[\frac{\partial \epsilon_r}{\partial y} \right]_{z=h} \frac{\partial h}{\partial y} \right).$$

Noting that

$$\int_0^h \frac{\partial}{\partial z} \left(\frac{\partial \ln \epsilon_r}{\partial y} \right) dz = \frac{1}{\epsilon_{ro}} \left[\frac{\partial \epsilon_r}{\partial y} \right]_{z=h}$$

and (using Leibniz's formula again)

$$\int_0^h \frac{\partial}{\partial y} \left(\frac{\partial \ln \epsilon_r}{\partial z} \right) dz = -\frac{1}{\epsilon_{ro}} \left[\frac{\partial \epsilon_r}{\partial z} \right]_{z=h} \frac{\partial h}{\partial y}$$

we have

$$\left[\frac{\partial \epsilon_r}{\partial y} \right]_{z=h} = - \left[\frac{\partial \epsilon_r}{\partial z} \right]_{z=h} \frac{\partial h}{\partial y}$$

since

$$\int_0^h \frac{\partial}{\partial z} \left(\frac{\partial}{\partial y} \ln \epsilon_r \right) dz = \int_0^h \frac{\partial}{\partial y} \left(\frac{\partial}{\partial z} \ln \epsilon_r \right) dz.$$

Hence, the HH object function becomes

$$O_{HH} = \frac{1}{16\pi^2 R^2} \exp(ik_0 x) \left[k_0^2 \gamma_0 h + \frac{1}{\epsilon_{ro}} \left[\frac{\partial \epsilon_r}{\partial z} \right]_{z=h} \left(\frac{\partial h}{\partial y} \right)^2 \right].$$

A relatively simple expression for the VV and HH RAR images can then be obtained by letting

$$\frac{1}{\epsilon_{ro}} \left[\frac{\partial \epsilon_r}{\partial z} \right]_{z=h} = k_0 \gamma_0 \simeq 1.8 \times 10^4 \mathrm{m}^{-1}.$$

Here it is assumed that the gradient in the vertical direction due to a change in the permittivity across the interface between the sea and air is equal to $k_0 \gamma_0 \epsilon_{ro} \simeq 1.3 \times 10^6 \mathrm{m}^{-1}$ over the imaged scene. This allows us to write the VV and HH RAR images as

$$I_{RAR}^{VV}(x,y) = A \mid \mathrm{sinc}(\alpha T x) \exp(-ik_0 y^2/R) \otimes\otimes \exp(ik_0 x) \left[1 + k_0 h(x,y) \right] \mid$$

and

$$I_{RAR}^{HH}(x,y)$$

$$= A \mid \mathrm{sinc}(\alpha T x) \exp(-ik_0 y^2/R) \otimes\otimes \exp(ik_0 x) \left[k_0 h(x,y) + \left(\frac{\partial}{\partial y} h(x,y) \right)^2 \right] \mid$$

where A is given by

$$A = \frac{\gamma_0 k_0 T}{16\pi^2 R^2} \simeq \frac{114T}{R^2}.$$

In this form, it is clear that the VV RAR image is a map of the height variations h of the sea surface whereas the HH RAR image is a map of both h and $(\partial_y h)^2$.

Compared to h, the nonlinear term $(\partial_y h)^2$ is very sensitive to the sea state. From this result we deduce that sea spikes are caused by rapid variations in the height of the sea surface as a function of the azimuth direction. In other words, the HH RAR image is dominated by features which map the location of points where

$$\left|\frac{\partial h}{\partial y}\right| >> k_0 h$$

on the scale of a wavelength. A simple illustration of this is given in Figure 18 which shows images of $\mid s_{ij} \mid$ and $\mid (s_{i(j+1)} - s_{ij})^2 \mid$, where s_{ij} is a 32×32 random Gaussian distributed array which is taken to represent a surface patch (without any deterministic patterns), each pixel being taken to be on the scale of a wavelength. A sequence of randomly distributed spikes occurs at locations where the difference between the $(j+1)^{th}$ and j^{th} elements of s_{ij} is relatively large so that the nonlinear term $(s_{i(j+1)} - s_{ij})^2$ produces a 'spike dominant' effect.

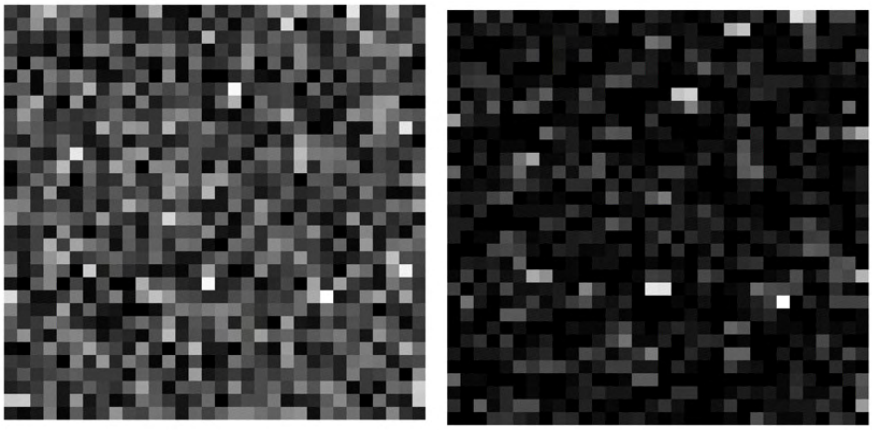

FIGURE 18 Simulation of sea spikes (right) using a low resolution rough surface patch model (left) for the sea surface.

5.4 Quantitative Imaging

The object functions for a SAR image show that the VV polarization data are related to both the permittivity and conductivity whereas the VH cross polarization data are related to the permittivity alone. This result provides a method of quantitative imaging using SAR. To illustrate the principle, consider a model where the grazing angle approaches zero and where the ground is composed of conductors embedded in a homogeneous dielectric. Using this model, we can employ the following conditions

$$\epsilon_r(x,y,z) = \epsilon_{ro}, \quad 0 \leq z \leq h(x,y)$$

and

$$\left[\frac{\partial \epsilon_r}{\partial x}\right]_{z<h} = \left[\frac{\partial \epsilon_r}{\partial y}\right]_{z<h} = \left[\frac{\partial \epsilon_r}{\partial z}\right]_{z<h} = 0.$$

The problem is then reduced to that of processing and combining the VV and VH polarization data in such a way that the reflections from conductors are isolated. As $\theta \to 0$, the SAR data for ij polarization are given by

$$D_{ij}(x,y) = P(x,y) \otimes\otimes O_{ij}(x,y)$$

where

$$O_{VV} = \frac{1}{16\pi^2 R^2} \exp(ik_0 x) \left(k_0^2 \gamma_0 h + \frac{1}{\epsilon_{ro}} \left[\frac{\partial \epsilon_r}{\partial z} \right]_{z=h} - ik_0 z_0 \int_0^h \sigma dz \right)$$

and

$$O_{VH} = \frac{1}{16\pi^2 R^2} \exp(ik_0 x) \frac{1}{\epsilon_{ro}} \left(\left[\frac{\partial \epsilon_r}{\partial z} \right]_{z=h} \frac{\partial h}{\partial y} \right).$$

If we then consider the case when

$$\frac{1}{\epsilon_{ro}} \left[\frac{\partial \epsilon_{ro}}{\partial z} \right]_{z=h} = k_0 \gamma_0$$

the object functions reduce to

$$O_{VV} = \frac{k_0 \gamma_0}{16\pi^2 R^2} \exp(ik_0 x) \left(1 + k_0 h - \frac{iz_0}{\gamma_0} \int_0^h \sigma dz \right)$$

and

$$O_{VH}(x,y) = \frac{k_0 \gamma_0}{16\pi^2 R^2} \exp(ik_0 x) \frac{\partial h}{\partial y}.$$

The VV and HH processed SAR data are then given as

$$D_{VV} = \frac{k_0 \gamma_0}{16\pi^2 R^2} P \otimes\otimes \exp(ik_0 x) \left(1 + k_0 h - \frac{iz_0}{\gamma_0} \int_0^h \sigma dz \right)$$

and

$$D_{VH} = \frac{\gamma_0}{16\pi^2 R^2} P \otimes\otimes \exp(ik_0 x) \frac{\partial}{\partial y} (1 + k_0 h)$$

where P is the point spread function. The last equation can be integrated directly giving

$$\frac{k_0 \gamma_0}{16\pi^2 R^2} P \otimes\otimes \exp(ik_0 x)(1 + k_0 h) = k_0 \int^y D_{VH} dy$$

and, hence, the VV polarization data can be written as

$$D_{VV} = k_0 \int^y D_{VH} dy - \frac{ik_0 z_0}{16\pi^2 R^2} P \otimes\otimes \exp(ik_0 x) \int_0^h \sigma dz.$$

By defining the SAR image of the conductivity variations as

$$I_{SAR}^{\sigma}(x,y) = \frac{k_0 \gamma_0}{16\pi^2 R^2} \mid P(x,y) \otimes\otimes \exp(ik_0 x) \int_0^{h(x,y)} \sigma(x,y,z) dz \mid$$

we then obtain

$$I^{\sigma}_{SAR}(x,y) = \mid D_{VV}(x,y) - k_0 \int^{y} D_{VH}(x,y)dy \mid .$$

This result provides a quantitative SAR image of the conductivity of the ground surface obtained by integrating the cross polarization data D_{VH}.

6 INVERSE SCATTERING SOLUTIONS WITH APPLICATIONS TO ELECTROMAGNETIC SIGNAL PROCESSING

6.1 Introduction

When a signal is recorded that has been physically generated by some scattering process (the interaction of electromagnetic inhomogeneous materials, for example), the 'standard model' for the signal (i.e. information content convolved with a characteristic Impulse Response Function) is usually based on a single scattering approximation. An additive noise term is introduced into the model to take into account a range of non-deterministic factors including multiple scattering that, along with electronic noise and other background noise sources, is assumed to be relatively weak. Thus, the standard model is based on a 'weak field condition' and the inverse scattering problem is often reduced to the deconvolution of a signal in the presence of additive noise.

Attempts at solving the exact inverse scattering problem for equations such as the inhomogeneous Schrödinger equation in quantum mechanics and the inhomogeneous Helmholtz equation in electromagnetism often prove to be intractable, particularly with regard to the goal of implementing algorithms that are computationally stable and/or compatible with standard signal analysis methods and Digital Signal Processing 'toolkits'. This chapter is concerned with an approach at solving the multiple scattering problem for narrow side-band systems (typically, electromagnetic signal processing systems) that is compounded in the introduction of a single extra term to the standard model. The approach is based on applying certain conditions to an exact solution of the inverse scattering problem rather than applying conditions to the forward scattering problem and then inverting the (conditional) result.

6.2 The Standard Model: Convolution Transform

The 'standard model' used to describe a signal $s(t)$ is based on the equation [23]

$$s(t) = p(t) \otimes f(t) + n(t)$$

where $\otimes$ denotes the convolution integral, $p(t)$ is the Impulse Response Function (IRF), $f(t)$ is the 'input' (representing the information content of the signal associated with some physical process - a scattering process, for example) and $n(t)$ is the characteristic noise of the system whose 'output' is $s(t)$. This linear and stationary model is based on a convolution transform and applies to a wide range of systems and applications involving the generation, interpretation and processing of digital signals and images.

In any application, a fundamental signal processing problem is to recover $f(t)$ from $s(t)$ given an estimate of $p(t)$ and $\mathbf{Pr}[n(t)]$ (where Pr denotes the Probability Density Function of the stochastic function $n(t)$). In many applications, this model can be taken to be a relatively complete and accurate representation of the physical processes that contribute to the generation of the signal $s(t)$, the noise function $n(t)$ being taken to be a combination of electronic and background noise associated with the recording system, for example. However, in applications where the signal is based on the scattering of an incident wavefield with an inhomogeneous medium, the standard model is based on applying an approximation to the scattering equation, e.g. the Helmholtz equation. The approximation is often referred to as the Born approximation (after Max Born, who first considered the approximation with regard to scattering processes in quantum mechanics through solutions to the Schrödinger equation) and considers the scattering events that contribute to the signal $s(t)$ to be based on single scattering processes only. This requires that the 'scattering model' adheres to the 'weak field' condition in which the total scattered field is considered to be a weak perturbation of the incident field in terms of some appropriate measure. In turn, depending on the complexity of the scattering model, this condition can usually be quantified in terms of physical parameters such as the wavelength λ of the incident wavefield and the scale length L of the scatterer, a basic 'standard' being that $\lambda >> L$. However, this condition is fundamentally incompatible with a basic requirement associated with systems that are designed to recover information at a resolution compatible with the scale of the wavelength, i.e. when $\lambda \sim L$. Thus, any system that is designed and engineered to 'image' an object on the scale of the wavelength of the incident field is prone to distortion due to the effect of multiple scattering, an effect that is not incorporated within the standard model. Instead, multiple scattering processes are considered to contribute to the noise function $n(t)$.

Given the standard model, the inverse (weak) scattering problem can be reduced to the process of Fourier inversion and deconvolution in the presence of additive noise. Ideally, we consider some function $p(t)$ such that $p(t) \odot p(t) = \delta(t)$ where $\odot$ denotes the correlation integral and $\delta(t)$ is the delta function, from which it follows that

$$f(t) = p(t) \odot s(t) + n'(t)$$

where $n'(t) = -p(t) \odot n(t)$.

In this chapter we consider an new approach to the inverse scattering problem that is *exact* in the sense that it includes the effects of multiple scattering [24]. Inverse solutions to the multiple scattering problem have been studied for many years and a variety of solutions developed. However, in practice, except for some special circumstances, such solutions are usually incompatible with the engineering of a system associated with the methods of signal analysis upon which it is based and the Digital Signal Processing 'toolkits' that are currently available, e.g. the MATLAB DSP 'Toolbox' [25]. In this chapter, we develop a model for the signal that is compounded in the equation

$$s(t) = p(t) \otimes [f(t) + \mid s(t) \mid^2 \exp(i\omega_0 t)] + n(t)$$

where the noise function $n(t)$ is not inclusive of the multiple scattering processes that are, instead, described by the term $p(t) \otimes [\mid s(t) \mid^2 \exp(i\omega_0 t)]$. This result is based on two basic conditions: (i) the total wavefield (i.e. the sum of the incident and scattered wavefields) is a phase only field; (ii) the frequency bandwidth is small compared to the carrier wave of the scattered field. In this sense, the result is not an exact inverse solution in itself but based on an exact inverse solution to which conditions (i) and (ii) are then applied. These conditions are applicable to narrow side-band pulse-echo systems, and, for example, side-band systems that exploit (linear) frequency modulated (FM) pulses. Side-band systems are a general characteristic of inverse problems associated with electromagnetic signals where the band-width is small compared with the carrier frequency. Such systems include Real Aperture Radar and Synthetic Aperture Radar and, in the latter case, a demonstration of the technique is provided by way of an application in electromagnetic signal processing.

This chapter is structured into the following sections: Section 6.3 provides a background to the forward and inverse scattering problems under the weak field (Born approximation for single scattering) in one-dimension, the weak gradient (WKB approximation) and strong field (multiple scattering) conditions. Section 6.4 introduces an exact inverse scattering approach and includes example numerical simulations for weak and strong scattering in one- and two-dimensions. The simulations considered are based on conditions appropriate for application to side-band signal processing systems where the bandwidth of the wavefield is significantly small compared to the carrier frequency which is taken to be high and Section 6.5 adopts the same approach for modelling narrow side-band pulse-echo systems. Section 6.5 addresses signal processing methods associated with FM pulse-echo systems which provides a background to the application of the solutions to Synthetic Aperture Radar imaging discussed in Chapter 5 as presented in Section 6.6.

The material presented in this paper is based on an analysis of the problem reduced to working in one-dimension. However, the approach is directly applicable to inverse scattering problems concerned with two-dimensional systems (e.g. diffraction tomography) and three-dimensional systems associated with radio, microwave, TeraHertz, photonics and electromagnetic signal and imaging processing systems in general.

6.3 Forward and Inverse Scattering Solutions in One-Dimension

Based on the material presented in Chapter 2, we consider a model in which ϵ and σ are one-dimensional functions, $\mu = \mu_0$ and the electric field is plane polarized, i.e. $\widetilde{\mathbf{E}} = \hat{\mathbf{z}}u(x,k)$, so that equation (8) can be reduced to the form (the one-dimensional inhomogeneous Helmholtz equation)

$$\left(\frac{\partial^2}{\partial x^2} + k^2\right) u(x,k) = -k^2\gamma(x,k)u(x,k)$$

where, for a conductive dielectric,

$$\gamma(x,k) = \gamma_\epsilon(x) - i\frac{z_0\sigma(x)}{k}$$

and for a non-conductive dielectric

$$\gamma(x) = \gamma_\epsilon(x) = \epsilon_r - 1$$

with $\epsilon = \epsilon_0\epsilon_r$ where $\epsilon_r > 1$ is the relative permittivity. The functions ϵ_r and σ are taken to be real whereas the wavefield u is a complex function with variations as a function of x and k in both amplitude $A_u(x,k)$ and phase $\theta_u(x,k)$, i.e.

$$u(x,k) = A_u(x,k)\exp[i\theta_u(x,k)]$$

On the basis of this model, the forward scattering problem is defined as: Given γ obtain a solution for u. The inverse scattering problem is: Given u derive a solution for γ.

We consider the case where the medium is a non-conductive dielectric and where $u(x,k),\ \ x \in (-\infty,\infty)$ is given by the sum of an incident wavefield $u_i(x,k)$ and a scattered wavefield $u_s(x,k)$, u_i being given by a solution of

$$\left(\frac{\partial^2}{\partial x^2} + k^2\right) u_i(x,k) = 0.$$

Thus, with $u = u_i + u_s$, for a non-conductive dielectric,

$$\left(\frac{\partial^2}{\partial x^2} + k^2\right) u_s(x,k) = -k^2\gamma(x)[u_i(x,k) + u_s(x,k)]. \tag{6.1}$$

For most practically significant cases, it may be assumed that γ is of compact support, i.e. $\gamma(x)\exists\forall x \in [-X,X]$.

6.3.1 Weak Field Condition and the Born Approximation

Given equation (6.1), the weak field condition is based on assuming that the contribution of the scattered field on the right hand side of this equation is minimal. Under this condition, equation (6.1) is, to a good approximation, given by

$$\left(\frac{\partial^2}{\partial x^2} + k^2\right) u_s(x,k) = -k^2\gamma(x)u_i(x,k),$$

provided

$$\|u_s(x,k)\| << \|u_i(x,k)\|$$

where $\|\bullet\|$ denotes the norm of the function over x. This is the Born approximation and provides a Green's function solution for the scattered field given by [23]

$$u_s(x,k) = k^2 g(\mid x \mid, k) \otimes \gamma(x) u_i(x,k)$$

where $\otimes$ denotes the convolution integral, i.e.

$$g(\mid x \mid, k) \otimes \gamma(x) u_i(x,k) \equiv \int_{-\infty}^{\infty} g(\mid x - y \mid, k) \gamma(y) u_i(y,k) dy.$$

Here, g is the 'outgoing' Green's function given by

$$g(\mid y - x \mid, k) = \frac{i}{2k} \exp(ik \mid y - x \mid)$$

which is the solution of

$$\left(\frac{\partial^2}{\partial x^2} + k^2 \right) g(\mid y - x \mid, k) = -\delta(y - x).$$

Formally, this solution requires that u and $\partial u / \partial x$ are zero at $x = \pm\infty$.

6.3.2 Asymptotic Solution

For the case when the wavefield is detected in the far field, i.e. $\mid x \mid >> \mid X \mid$, we can consider an asymptotic solution of the form

$$u_s(x,k) = \lim_{x \to \infty} \frac{ik}{2} \int_{-X}^{X} \exp(ik \mid x - y \mid) \gamma(y) u_i(y,k) dy$$

$$= \exp(ikx) \frac{ik}{2} \int_{-X}^{X} \exp(-iky) \gamma(y) u_i(y,k) dy.$$

For an incident field $u_i(x,k) = \exp(-ikx)$, $u = u_i + u_s$ is now given by

$$u(x,k) = u_i(x,k) + u_s(x,k) = \exp(-ikx) + \widetilde{s}(k) \exp(ikx)$$

where

$$\widetilde{s}(k) = \frac{ik}{2} \widetilde{\gamma}(k),$$

$$\widetilde{\gamma}(k) = \int_{-\infty}^{\infty} \gamma(x) \exp(-2ikx) dx.$$

This solution for u represents the right- and left-travelling components of the wavefield, the latter case being determined by the 'reflection coefficient' $\widetilde{s}(k)$. Note that in this asymptotic solution, the function $\gamma(x)$ maps to its Fourier transform $\widetilde{\gamma}(k)$ (ignoring scaling). Thus, at a fixed position in space $\mid x \mid >> \mid X \mid$, the function $\gamma(x)$ can only be recovered from information on its spectrum $\widetilde{\gamma}(k)$. In this sense, the inverse scattering problem is reduced to the problem of Fourier inversion. In practice, this requirement necessitates the application of pulse-echo methods which are discussed later.

6.3.3 The Weak Gradient (WKB) Approximation

The solutions considered so far have been based on the application of the Born approximation (Born scattering theory) to Green function solutions of time-independent wave equations. In this section, we consider the Wentzel-Kramers-Brillouin (WKB) which is similar to the Rytov approximations for solving the one-dimensional inhomogeneous Helmholtz equations,

The WKB method is based on the idea that if the wavelength of the wavefield u is very small compared to variations in γ then a suitable approximation can be introduced which provides an appropriate solution. The approach is based on the use of an exponential type or 'Eikonal' transformation where a solution of the form $A(x,k)\exp[\pm s(x,k)]$ is considered with amplitude function $A(x,k)$ and phase function $s(x,k)$ [26]. This is analogous to a plane wave solution of the type $A\exp(\pm ikx)$. A historically important example of the WKB approximation being used was in a paper by George Green on *The Motion of Waves in a Variable Canal of Small Depth and Width* (published in the Transactions of the Cambridge Philosophical Society in 1837) who developed a solution for waves along a narrow (to make the problem one dimensional) but variable channel. His solution involves an approach which is essentially the same as the WKB method used in quantum mechanics. It is therefore arguable that the approximation should be called the Green approximation!

To illustrate the idea behind the WKB approximation, let us consider a general solution to the 1D wave equation

$$\left(\frac{\partial^2}{\partial x^2}+k^2\right)u(x,k)=-k^2\gamma(x)u(x,k). \tag{3.4}$$

The Green function solution to this equation is given by

$$u=u_i+u_s$$

where u_i is the incident wavefield (typically a unit amplitude plane wave) and u_s is given by

$$u_s(x_0,k)=k^2\int\gamma(x)g(x\mid x_0,k)u(x,k)dx.$$

Instead of considering the solution to be the sum of two wavefields u_i and u_s, suppose we introduce the eikonal transform

$$u(x,k)=u_i(x,k)\exp[s(x,k)].$$

Substituting this result into equation (6.2) and differentiating, we obtain

$$\frac{\partial^2 u_i}{\partial x^2}+2\frac{\partial s}{\partial x}\frac{\partial u_i}{\partial x}+u_i\left(\frac{\partial s}{\partial x}\right)^2+u_i\frac{\partial^2 s}{\partial x^2}+k^2u_i=-k^2\gamma u_i.$$

If we now consider u_i to be a solution to $\partial^2 u_i/\partial x^2+k^2u_i=0$ then, after differentiating u_i and rearranging, we have

$$2ik\frac{\partial s}{\partial x}+\left(\frac{\partial s}{\partial x}\right)^2+\frac{\partial^2 s}{\partial x^2}=-k^2\gamma. \tag{6.3}$$

This is a nonlinear Riccatian equation for s which at first sight, appears to be more complicated than the original. However, if we introduce the condition that the wavelength $\lambda = 2\pi/k$ is significantly smaller than the spatial extent over which s varies, then the nonlinear term and the second derivative can be ignored and we can write

$$2ik\frac{ds}{dx} = -k^2\gamma$$

whose general solution is (ignoring the constant of integration)

$$s(x) = \frac{ik}{2}\int^x \gamma(x)dx.$$

The solution for u is therefore given by

$$u(x,k) = u_i(x,k)\exp\left(\frac{ik}{2}\int^x \gamma(x)dx\right)$$

$$= \exp\left[ik\left(x + \frac{1}{2}\int^x \gamma(x)dx\right)\right].$$

This is an example of the WKB approximation. It is based on the idea that if k is large compared to the magnitudes of the terms $(\partial s/\partial x)^2$ and $\partial^2 s/\partial x^2$ then the only terms in equation (6.3) that matter are $2ik(\partial s/\partial x)$ and $-k^2\gamma$. In other words, if L is the characteristic scale length over which s varies, then

$$\frac{\lambda}{L} << 1.$$

The solution describes a plane wavefield whose phase kx is modified by $\frac{k}{2}\int \gamma dx$, the inverse scattering solution being given by (ignoring scaling) [27]

$$\gamma(x) \sim \frac{d}{dx}\ln\left[\frac{u(x,k)}{u_i(x,k)}\right].$$

A similar approach can be used in higher dimensions which leads to an interpretation of the solutions in terms of the characteristics or rays and the geometric properties associated with them.

The WKB approximation as illustrated here does not in itself make use of a Green function whereas the Rytov approximation discussed in Chapter 3 is based on a similar idea to the WKB approximation and makes explicit use of the Green function.

6.3.4 Solution for Multiple Scattering

The Born approximation can be considered to be a first solution u_1 to the iterative series (for $n = 1, 2, 3, ...$)

$$u_{n+1}(x,k) = u_i(x,k) + k^2 g(\mid x \mid, k) \otimes \gamma(x)u_n(x,k)$$

where $u_0 = u_i$. The scattered field, can be written in the form (Born series solution)

$$u_s(x,k) = k^2 g(\mid x \mid, k) \otimes \gamma(x) u_i(x,k)$$

$$+k^4 g(\mid x \mid, k) \otimes \gamma(x)[g(\mid x \mid, k) \otimes \gamma(x) u_i(x,k)] + ...$$

where each term in this series expresses the effects due to single, double and triple etc. scattering, i.e. the wavefields generated by an increasing number of interactions.

In principle, if this series converges, then it must converge to the solution. Using operator notation and writing

$$u_{n+1} = u_i + \hat{I} u_n$$

where

$$\hat{I} = k^2 \int dx g \gamma,$$

at each iteration n we consider the solution to be given by

$$u_n = u + \epsilon_n$$

where ϵ_n is the error associated with the solution at iteration n and u is the exact solution. A necessary condition for convergence is then $\epsilon_n \to 0$ as $n \to \infty$. Since

$$u + \epsilon_{n+1} = u_i + \hat{I}(u + \epsilon_n) = u_i + \hat{I} u + \hat{I} \epsilon_n$$

we can write

$$\epsilon_{n+1} = \hat{I} \epsilon_n$$

given that $u = u_i + \hat{I} u$. Thus

$$\epsilon_1 = \hat{I} \epsilon_0; \;\; \epsilon_2 = \hat{I} \epsilon_1 = \hat{I}(\hat{I} \epsilon_0); \;\; \epsilon_3 = \hat{I} \epsilon_2 = \hat{I}[\hat{I}(\hat{I} \epsilon_0)]; \;\; ...$$

or

$$\epsilon_n = \hat{I}^n \epsilon_0$$

from which it follows that

$$\| \epsilon_n \| = \| \hat{I}^n \epsilon_0 \| \leq \| \hat{I}^n \| \times \| \epsilon_0 \| \leq \| \hat{I} \|^n \| \epsilon_0 \|.$$

The condition for convergence therefore becomes

$$\lim_{n \to \infty} \| \hat{I} \|^n = 0.$$

This is only possible if

$$\| \hat{I} \| < 1$$

or

$$k^2 \| g(\mid x \mid, k) \otimes \gamma(x) \| < 1.$$

Defining the Euclidean norm of a complex function $f(x)$ to be

$$\| f(x) \|_2 = \left(\int \mid f(x) \mid^2 dx \right)^{\frac{1}{2}}$$

we have

$$k^2\|g(\mid x \mid,k)\otimes\gamma(x)\|_2 \leq k^2\|g(\mid x \mid,k)\|_2\|\gamma(x)\|_2$$

and noting that $x \in [-X, X]$, we can write

$$kX\langle\gamma\rangle < 1$$

where

$$\langle\gamma\rangle \equiv \left(\int_{-X}^{X} \mid \gamma(x) \mid^2 dx\right)^{\frac{1}{2}}.$$

This is the condition required for the Born series to converge, the Born approximation being dependent on the weak field condition

$$\langle\gamma\rangle << \frac{\lambda}{X}.$$

6.3.5 Inverse Solution for Multiple Scattering Processes

Using operator notation, the Born series can be written as

$$u(x,k) = u_i(x,k) + \hat{I}_i\gamma(x,k) + \hat{I}_i(\gamma(x,k)\hat{I}\gamma(x,k))$$
$$+I_i[\gamma(x,k)\hat{I}(\gamma(x,k)\hat{I}\gamma(x,k))] + ...$$

where $\gamma(x,k) = k^2\gamma(x)$ and

$$\hat{I}_i = \int dx u_i g, \quad \hat{I} = \int dx g.$$

Let $\epsilon U = u - u_i$ and

$$\gamma = \sum_{j=1}^{\infty} \epsilon^j \gamma_j.$$

Then

$$\epsilon U = \hat{I}_i[\epsilon\gamma_1 + \epsilon^2\gamma_2 + \epsilon^3\gamma_3 + ...]$$
$$+\hat{I}_i[(\epsilon\gamma_1 + \epsilon^2\gamma_2 + \epsilon^3\gamma_3 + ...)\hat{I}(\epsilon\gamma_1 + \epsilon^2\gamma_2 + \epsilon^3\gamma_3 + ...)]$$
$$+\hat{I}_i\{(\epsilon\gamma_1 + \epsilon^2\gamma_2 + \epsilon^3\gamma_3 + ...)\hat{I}[(\epsilon\gamma_1 + \epsilon^2\gamma_2 + \epsilon^3\gamma_3 + ...)$$
$$\hat{I}(\epsilon\gamma_1 + \epsilon^2\gamma_2 + \epsilon^3\gamma_3 + ...)]\} + ...$$

Equating terms with common coefficients ϵ, ϵ^2 etc. we have For $j = 1$:

$$U = \hat{I}_i\gamma_1; \quad \gamma_1 = \hat{I}_i^{-1}U.$$

For $j = 2$:

$$0 = \hat{I}_i\gamma_2 + \hat{I}_i(\gamma_1\hat{I}\gamma_1); \quad \gamma_2 = -\hat{I}_i^{-1}[\hat{I}_i(\gamma_1\hat{I}\gamma_1)]$$

and so on. By computing the functions γ_j using this iterative method, the scattering function γ is obtained by summing γ_j for $\epsilon = 1$. This approach provides a formal exact inverse scattering solution [28], [29] but it is not unconditional, i.e. the inverse solution is only applicable when the Born series converges to the exact scattering solution and thus when

$$\|g(\mid x \mid,k)\otimes\gamma(x)\| < 1.$$

Note that for $j = 1$, the solution for γ_1 is that obtained under the Born approximation.

6.4 Exact Inverse Scattering Solutions

Given equation (6.2), an exact inverse scattering solution can be formulated based on defining γ as

$$\gamma(x) = \frac{u^*(x,k)}{\mid u(x,k) \mid^2} \frac{\partial^2}{\partial x^2} \left[R(x) \otimes u_s(x,k) - \frac{1}{k^2} u_s(x,k) \right].$$

where (c_1 and c_2 being arbitrary constants)

$$R(x) = \begin{cases} (c_1 - 1)x + c_2, & x > 0; \\ 0, & \text{otherwise.} \end{cases}$$

This result is derived in Appendix 1 and is an extension of the trivial solution

$$\gamma(x) = -\frac{1}{k^2 u(x,k)} \left(k^2 + \frac{\partial^2}{\partial x^2} \right) u(x,k)$$

which, noting that $\gamma = \epsilon_r - 1$ can be written in the form

$$\epsilon_r(x) = -\frac{u^*(x,k)}{k^2 \mid u(x,k) \mid^2} \frac{\partial^2}{\partial x^2} u(x,k).$$

However, since $u = u_i + u_s$ where u_i is a solution of

$$\left(\frac{\partial^2}{\partial x^2} + k^2 \right) u_i(x,k) = 0 \tag{6.4}$$

we can write

$$\gamma(x) = -\frac{u_i^*(x,k) + u_s^*(x,k)}{k^2 \mid u_i(x,k) + u_s(x,k) \mid^2} \left(k^2 + \frac{\partial^2}{\partial x^2} \right) u_s(x,k).$$

Further, we note that under the weak field condition where we consider the equation

$$\left(\frac{\partial^2}{\partial x^2} + k^2 \right) u_s(x,k) = -k^2 \gamma(x) u_i(x,k),$$

the equivalent expression for γ (under the Born approximation) is

$$\gamma(x) = -\frac{u_i^*(x,k)}{k^2 \mid u_i(x,k) \mid^2} \left(k^2 + \frac{\partial^2}{\partial x^2} \right) u_s(x,k)$$

$$= -\frac{u_i^*(x,k)}{k^2} \left(k^2 + \frac{\partial^2}{\partial x^2} \right) u_s(x,k)$$

given that $u_i(x,k) = \exp(ikx)$ is a solution of equation (6.4).

6.4.1 Narrow Side-band Condition

In general, and in comparison to the solution under the Born approximation, the exact inverse scattering solution for γ includes $\mid u(x,k) \mid^{-2}$ which describes the (inverse square) amplitude envelop of the sum of the incident and scattered fields, i.e. with $u(x,k) = A_u(x,k)\exp[i\theta_u(x,k)]$, it is clear that $\mid u(x,k) \mid^2 = [A_u(x,k)]^2$. If u is taken to be a narrow side-band signal that is dominated by a high frequency carrier wave determined by k_0, which, in turn, is determined by a narrow band incident wave u_i, then, we can consider the condition where $\mid u(x,k) \mid^2 \sim 1 \forall x$ (in general, a constant). This condition is equivalent to considering the case where u is a phase only field $u(x,k) = \exp[i\theta_u(x,k)]$ say, and is imposed in respect of the fact that, for narrow side-band signals, the contribution of $\mid u \mid^{-2}$ to the reconstruction of γ from u is insignificant when compared to $u^*(k_0^2 + \partial_x^2)u_s$.

Strictly speaking, the condition being imposed implies that, with $u_i = \exp(\pm ik_0 x)$,

$$u_i^* u_s + (u_i^* u_s)^* + \mid u_s \mid^2 = 0.$$

The principal difference between the reconstruction of γ using the Born approximation and the inverse solution considered here is compounded in the addition of a single term, i.e. for a weak field where $\|u_s\| << \|u_i\|$

$$\gamma(x) = -u_i^*(x,k_0)\left(1 + \frac{1}{k_0^2}\frac{\partial^2}{\partial x^2}\right)u_s(x,k_0) \tag{6.5}$$

and for a strong field where $\|u_s\| \sim \|u_i\|$

$$\gamma(x) = -u_i^*(x,k_0)\left(1 + \frac{1}{k_0^2}\frac{\partial^2}{\partial x^2}\right)u_s(x,k_0)$$

$$-u_s^*(x,k_0)\left(1 + \frac{1}{k_0^2}\frac{\partial^2}{\partial x^2}\right)u_s(x,k_0). \tag{6.6}$$

Note that under the strict definition of a phase only field, for $\|u_s\| \sim \|u_i\|$

$$\gamma(x) = u_i(x,k_0)u_s^*(x,k_0)$$

$$-\frac{[u_i(x,k_0) + u_s(x,k_0)]^*}{k_0^2}\frac{\partial^2}{\partial x^2}u_s(x,k_0)$$

$$= -u_i(x,k_0)u_s^*(x,k_0), \quad k_0 \to \infty$$

which yield the weak field condition. In this paper, we relax this condition and utilze equations (6.5) and (6.6) given that $\mid u_i + u_s \mid^2 \sim 1$. This is the 'key' to the results that follow.

6.4.2 Numerical Simulation I: One-dimensional Model

We consider a numerical simulation based on computing the scattered field using a forward differencing approach to equation (6.2). For a fixed wavenumber k_0 (which defines a continuous wave mode), the vector u_n computed over a uniformly sampled discrete array composed of N line elements, each of length Δ, is given by

$$\frac{u_{n+1} - 2u_n + u_{n-1}}{\Delta^2} + k_0^2 u_n = -k_0^2 \gamma_n u_n, \;\; n \in [1, N]$$

which has the solution

$$u_n = \frac{u_{n+1} + u_{n-1}}{2 - \Delta^2 k_0^2 (1 + \gamma_n)} = \frac{u_{n+1} + u_{n-1}}{2 - \Delta^2 k_0^2 \epsilon_{r,n}}. \tag{6.7}$$

This solution requires the following iteration to be implemented

$$u_n^{m+1} = \frac{u_{n+1}^m + u_{n-1}^m}{2 - \Delta^2 k_0^2 \epsilon_{r,n}} \tag{6.8}$$

where u_n^1 is taken to be a discrete representation of the (unit amplitude) incident field which we define with a fixed number of periods p over $n \in [1, N]$, i.e.

$$u_n^1 = \exp\left(\frac{2\pi i p (n-1)}{(N-1)}\right).$$

A necessary condition that must be applied to equation (6.8) is that

$$\Delta k_0 < \sqrt{\frac{2}{\|\epsilon_{r,n}\|_\infty}}$$

where $\| \bullet \|_\infty$ defines the 'uniform norm' and

$$k_0 = \frac{2\pi p}{N}.$$

Since a solution to equation (6.2) is not necessarily conditional on the amplitude of the wavefield (i.e. the incident field can be $A \exp(\pm i k_0 x)$ where A is an arbitrary value), the amplitude of the array u_n^m, after M iterations is normalised on output, i.e. $u_n^M \to u_n^M / \|u_n^M\|_\infty$.

For the discrete case, equations (6.5) and (6.6) transform to (for M iterations used to compute the scattered field)

$$\epsilon_{r,n} = 1 - u_{i,n}^* [u_{s,n} + (\Delta k_0)^{-2} u_{s,n} \otimes (1, -2, 1)]$$

$$= 1 - (u_n^1)^* [(u_n^M - u_n^1) + (\Delta k_0)^{-2} (u_n^M - u_n^1) \otimes (1, -2, 1)] \tag{6.9}$$

and

$$\epsilon_{r,n} = 1 - u_n^* [u_{s,n} + (\Delta k_0)^{-2} u_{s,n} \otimes (1, -2, 1)]$$

$$= 1 - (u_n^M)^* [(u_n^M - u_n^1) + (\Delta k_0)^{-2} (u_n^M - u_n^1) \otimes (1, -2, 1)] \tag{18}$$

respectively where $\otimes$ now denotes the discrete convolution sum. Figures 19 to 21 show comparisons between the numerical results given by equations (6.9) and (18), illustrating the superiority of the inverse scattering solution considered over that given by the weak field solution for the case when $\Delta k_0 = 0.01$. These results are based on applying the initial solution $u_n^1 = \sin[2\pi ip(n-1)/(N-1)]$ and Hilbert transforming the output arrays before computation of equations (6.9) and (18) through use of the MATLAB function *hilbert*. The profile of the reconstruction for $\epsilon_{r,n}$ based on equation (18) is preserved inclusive of characteristic 'ringing' due to the discontinuities associated with the model for $\epsilon_{r,n}$. In comparison, the weak field solution given by equation (6.9) has a relatively narrow dynamic range and provides a poor reconstruction particularly with regard to resolving the discontinuities associated with $\epsilon_{r,n}$.

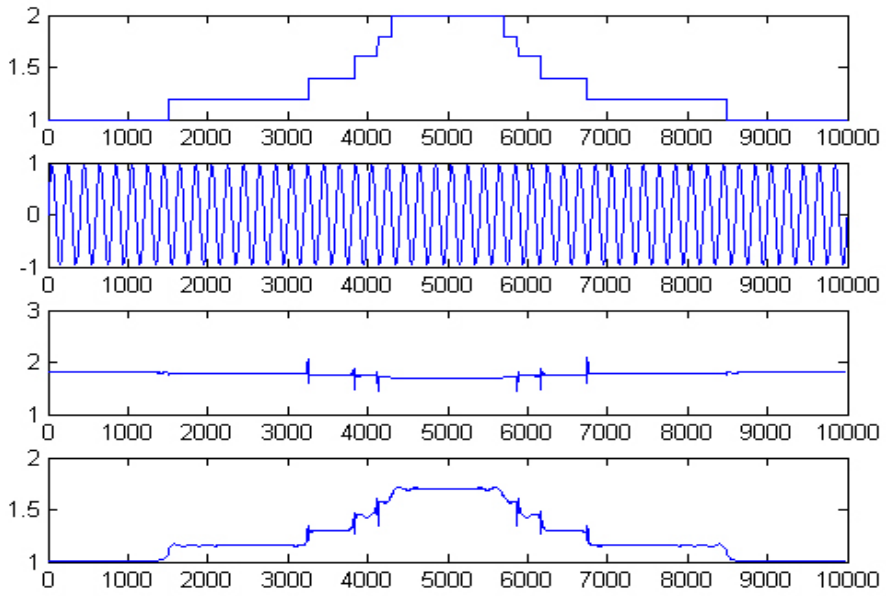

FIGURE 19 Comparison between the weak and strong field inverse scattering solutions for the case when $N = 10000$, $M = 100$, $\Delta k_0 = 0.01$ with $p = 50$. From top to bottom: Relative permittivity function model $\epsilon_{r,n}$; real part of wavefield computed via equation (6.8); inverse solution (real part) computed using equation (6.9); inverse scattering solution (real part) computed using equation (18).

6.4.3 Numerical Simulation II: Two-dimensional Model

Further appreciation of the difference between these weak and strong field solutions is realised in Figure 22. Here, we have considered an iterative forward scattering solution to the equation

$$\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + k_0^2\right) u(x,y,k_0) = -k_0^2 \gamma(x) u(x,y,k_0)$$

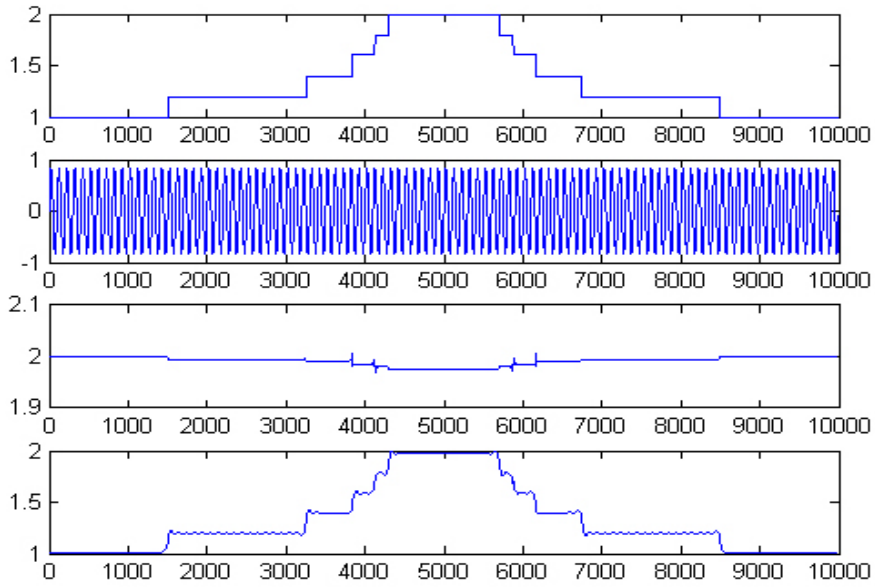

FIGURE 20 Comparison between the weak and strong field solutions for the case when $N = 10000$, $M = 100$, $\Delta k_0 = 0.01$ with $p = 100$. The descriptions of each plot follow those as given in Figure 19.

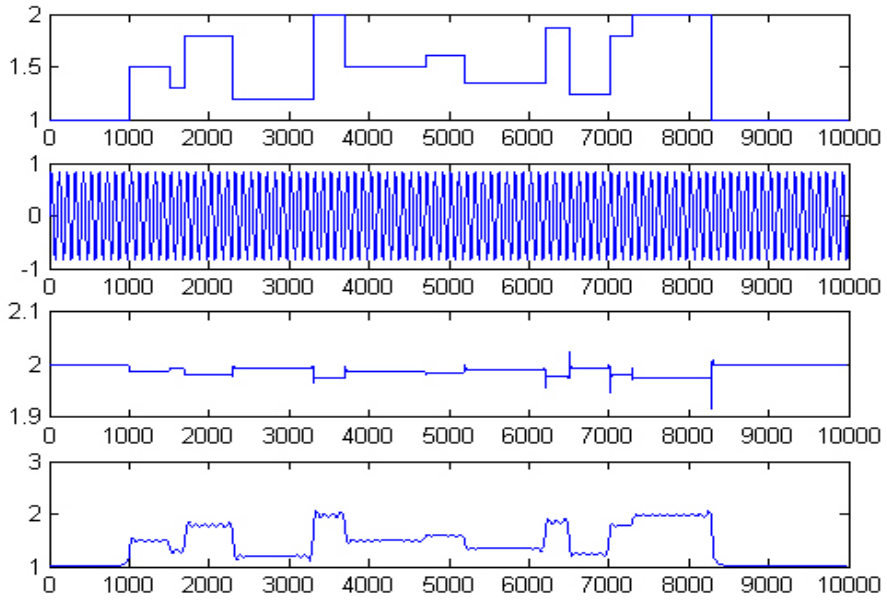

FIGURE 21 Comparison between the weak and strong field inverse scattering solutions for the case when $N = 10000$, $M = 100$, $\Delta k_0 = 0.01$ with $p = 50$ and a non-symmetric model of the relative permittivity $\epsilon_{r,n}$. The descriptions of each plot follow the same as those given in Figure 19.

.

based on the application of a regular (square) grid of size N^2, i.e.

$$\frac{u_{n+1,m} + u_{n-1,m} + u_{n,m+1} + u_{n,m-1} - 4u_{n,m}}{\Delta^2} + k_0^2 u_{n,m}$$

$$= -k_0^2 \gamma_{n,m} u_{n,m}, \quad n \in [1, N], \quad m \in [1, N]$$

with iterative solution

$$u_{n,m}^{q+1} = \frac{u_{n+1,m}^q + u_{n-1,m}^q + u_{n,m+1}^q + u_{n,m-1}^q}{4 - \Delta^2 k_0^2 \epsilon_{r,n,m}}$$

under the condition that

$$\Delta k_0 < \frac{2}{\sqrt{\| \epsilon_{r,n,m} \|_\infty}}.$$

with initial condition

$$u_{n,m}^1 = \sin[2\pi i p n (m-1)/(N-1)]$$

After M iteration, $u_{n,m}^M$ is normalised and the Hilbert transforms are taken of $u_{n,m}^M$, $u_{n,m}^1$ and $u_{n,m}^M - u_{n,m}^1$ over m for all values of n (to compute analytic signals $\forall n$) prior to the computation of the weak field reconstruction

$$\epsilon_{r,n,m} = 1 - (u_{n,m}^1)^* (u_{n,m}^M - u_{n,m}^1) - (\Delta k_0)^{-2} (u_{n,m}^1)^*$$

$$\times \left[(u_{n,m}^M - u_{n,m}^1) \otimes_2 \begin{pmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{pmatrix} \right] \tag{19}$$

and the strong field reconstruction

$$\epsilon_{r,n,m} = 1 - (u_{n,m}^M)^* (u_{n,m}^M - u_{n,m}^1) - (\Delta k_0)^{-2} (u_{n,m}^M)^*$$

$$\times \left[(u_{n,m}^M - u_{n,m}^1) \otimes_2 \begin{pmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{pmatrix} \right] \tag{20}$$

where $\otimes_2$ denotes the two-dimensional convolution sum.

6.5 Pulse-Echo Mode Signals

The example numerical results presented in the previous section have been introduced to illustrate the characteristic differences between the weak and strong field solutions given the exact inverse scattering approach that we have considered. However, in practice, these inverse solutions have little practical value with regard to engineering a system designed to recover $\epsilon_r(x)$ from information on the scattered field measured 'outside' the scatterer, i.e. when $\mid x \mid > \mid X \mid$. This is because the solutions considered so far require that the scattered field is known for $x \in [-X, X]$.

FIGURE 22 Comparison between the two-dimensional weak field and strong field inverse scattering solutions for the case when $N = 500$, $M = 100$, $\Delta k_0 = 0.01$ with $p = 64$ for the (left-to-right) CW case. Left: graded point scattering model for permittivity function $1 < \epsilon_{r,n,m} < 2$; centre: inverse solution (absolute value) computed using equation (19); right: inverse scattering solution (absolute value) computed using equation (20). Note that in each case, the numerical fields have been normalised for the purpose of generating grey level image displays.

Practically applicable systems typically measure the scattered field at a fixed point x_0 in the far field (based on an asymptotic solution where $x \to \infty$) by recording the spectrum $u_s(x_0, k)$ at x_0 over a range of values of k. Theoretically, this requires that the inverse problem is based on an asymptotic solution of the type derived in Section 6.3.2. In practice, this requires the application of pulse-echo mode methods.

Pulse-echo methods involve the emission of a pulse and a recording of the back-scattered wavefield (echo). This approach is consistent with the physical nature a system if: (i) the scattering function is a one-dimensional function; (ii) the incident wavefield is a 'pencil-line beam'. However, since all physical systems are intrinsically three-dimensional, this model is idealised. Nevertheless, a variety of electromagnetic information and imaging systems can be 'cast' in terms of problems involving layered materials (e.g. the response of light, radio and microwaves to layered dielectric materials including the propagation of electromagnetic waves along transmission lines such as an optical fiber).

Pulse-echo mode signal analysis systems typically involve the utilization a pulse that is emitted from a source in which the 'time history' of the back-scattered field is recorded by a receiver which is placed in the vicinity of the location of the source. By moving both the source and receiver and repeating this type of experiment, an image can be built up based on the nature of the reflected pulse at different source locations. The resolution that can be obtained with pulse-echo experiments of this type is normally determined by the length of the pulse that is used and the width of the beam. To obtain high resolution, a short pulse and narrow 'pencil beam' are required. In some cases, the lateral resolution can be synthesized using synthetic aperture methods. Also, in some cases (e.g. Real and Synthetic Aperture Radar), it is possible to modulate the frequency of the pulse thereby providing a method of reconstruction in which the resolution improves with pulse length (as discussed in Chapter 4).

In a pulse-echo experiment, the receiver monitors the time history of the reflected waves (the echo). After a short delay (which depends on the distance of the source from the scatterer and the speed at which the pulse propagates), the first reflections are received followed by a series or 'train' of other reflections from the surface or the interior of the material. This process continues until all the energy of the pulse has been dissipated. In electromagnetic systems (coherent time-history resolved), the scattered electric field is typically measured by the way in which it induces a time varying voltage in the antenna.

6.5.1 Base-band and Side-band Systems

Pulse-echo systems are based on using wavefields at frequencies where the time variations of the wavefield can be recorded to produce a set of signals. With electromagnetic systems, the frequency range is from kHz - GHz. Apart from synthetic aperture imaging systems, most pulse-echo based systems provide partially coherent (in time) data. There is one important difference between them, however, which is concerned with whether or not the pulse is a side-band of base-band wavefield. Baseband pulses are multi-frequency wavefields with a frequency range from 0-Ω Hz where Ω is the bandwidth of the pulse. Sideband pulses are fields with a bandwidth of Ω but with a central frequency of ω_0 (the carrier frequency of the pulse) where $\omega_0 >> \Omega$. In side-band systems, it is usual to demodulate back to base-band and then digitize the resulting signal(s). Sideband systems are a natural consequence of utilizing high frequency radiation sources where the pulse length is much longer than the wavelength. Thus, suppose a pulse of radiation denoted by $p(t)$ has a spectrum $\widetilde{p}(\omega)$ where $\mid \omega \mid \leq \Omega$. Then, for a base-band system we have

$$p(t) \leftrightarrow \widetilde{p}(\omega)$$

but for a side-band system

$$p(t)\exp(-i\omega_0 t) \leftrightarrow \widetilde{p}(\omega) \otimes \delta(\omega + \omega_0) = \widetilde{p}(\omega + \omega_0)$$

where $\leftrightarrow$ denotes the transformation from time to frequency space. In the latter case, there are many oscillations of the field over the duration of the pulse and hence we have $p(t)\exp(-i\omega_0 t)$ rather than just $p(t)$. Under the Born approximation, noting that $\omega = k/c_0$, for an incident field $u_i(t,\omega) = \exp(-i\omega t)$ the reflection coefficient - as derived in Section 6.3.2 - is

$$\widetilde{s}(\omega) = \frac{i\omega}{2c_0} \int_{-\infty}^{\infty} \gamma(t)\exp(-2i\omega t)dt.$$

where

$$\widetilde{\gamma}(\omega) = \int_{-\infty}^{\infty} \gamma(t/2)\exp(-i\omega t)dt.$$

Thus, for an incident field with a spectrum given by $\widetilde{p}(\omega)$, i.e. an incident field given by $u_i(t,\omega) = \widetilde{p}(\omega)\exp(-i\omega t)$, it follows that the reflection coefficient is given by

$$\widetilde{s}(\omega) = \frac{i\omega}{4c_0}\widetilde{p}(\omega)\widetilde{\gamma}(\omega)$$

or using the convolution theorem,

$$s(t) = p(t) \otimes f(t)$$

where

$$p(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \tilde{p}(\omega) \exp(i\omega t) d\omega$$

and $f(t)$ is the Impulse Response Function (IRF) given by (where t is the two-way travel time and we ignore scaling by $1/4c_0$)

$$f(t) = \frac{d}{dt}\gamma(t).$$

However, for a side-band band system

$$\tilde{s}(\omega) = i\omega\tilde{p}(\omega - \omega_0)\tilde{\gamma}(\omega) \simeq i\omega_0\tilde{p}(\omega)\tilde{\gamma}(\omega + \omega_0), \;\; \Omega << \omega_0$$

so that

$$s(t) = p(t) \otimes f(t)$$

where

$$f(t) = i\omega_0\gamma(t) \exp(-i\omega_0 t).$$

For a conductive dielectric, the (equivalent) IRFs are given by

$$f(t) = \frac{d}{dt}\gamma(t) + z_0\sigma(t)$$

and

$$f(t) = [i\omega_0\gamma(t) + z_0\sigma(t)] \exp(-i\omega_0 t)$$

for a base-band and side-band system respectively. Irrespective of whether a base-band or side-band pulse is used, the signal model

$$s(t) = p(t) \otimes f(t)$$

is based on the weak field condition. The 'standard' approach is to extended this model to the form

$$s(t) = p(t) \otimes f(t) + n(t)$$

where the noise term $n(t)$ is taken to include all effects that do not conform to the model used, including multiple scattering processes. The inverse scattering problem is thus reduced to the problem of deconvolution in the presence of additive noise. This is a fundamental problem in both signal and image processing.

6.5.2 Inverse Solution for Side-Band Pulse-Echo Systems

For a narrow side-band system with carrier frequency $k_0 >> \mid k \mid$, we consider the equation

$$\gamma(x) = -\frac{u^*(x,k)}{k_0^2}\left(k_0^2 + \frac{\partial^2}{\partial x^2}\right) u_s(x,k)$$

which, for a weak field where $\|u_s\| << \|u_i\|$ reduces to

$$\gamma(x) = -\frac{u_i^*(x,k)}{k_0^2}\left(k_0^2 + \frac{\partial^2}{\partial x^2}\right) u_s(x,k).$$

These results apply for all values of $x \in (-\infty, \infty)$ but γ is, as usual, considered to be of compact support $\gamma(x) \exists \forall x \in [-X, X]$. Under the Born approximation, for $x \to \infty$, a mapping is obtained between the scattered field u_s and γ that is based on the Fourier transform $\widetilde{\gamma}(k)$ of $\gamma(x)$, i.e.

$$u_s(x,k) = \frac{ik_0}{2}\exp(ik_0 x)\widetilde{\gamma}(k), \quad \mid k \mid << k_0$$

which is a solution of

$$\gamma(x) = -\frac{u_i^*(x,k)}{k_0^2}\left(k_0^2 + \frac{\partial^2}{\partial x^2}\right) u_s(x,k).$$

This result suggests taking the Fourier transform of γ where

$$\gamma(x) =$$

$$-[u_i^*(x,k_0) + u_s^*(x,k_0)]\left(1 + \frac{1}{k_0^2}\frac{\partial^2}{\partial x^2}\right) u_s(x,k_0)$$

giving

$$\widetilde{\gamma}(k) =$$

$$-[\widetilde{u}_i^*(k,k_0) + \widetilde{u}_s^*(k,k_0)] \odot \left(1 - \frac{\mid k \mid^2}{k_0^2}\right) \widetilde{u}_s(k,k_0)$$

$$= -[\widetilde{u}_i^*(k,k_0) + \widetilde{u}_s^*(k,k_0)] \odot \widetilde{u}_s(k), \quad \mid k \mid << k_0$$

where $\widetilde{\gamma}$, $\widetilde{u}_i^*$, $\widetilde{u}_s^*$ and $\widetilde{u}_s$ are the Fourier transforms of γ, u_i^*, u_s^* and u_s respectively and $\odot$ denotes the correlation integral. Noting that, for $u_i(x,k_0) = \exp(-ik_0 x)$, $\widetilde{u}_i^*(k,k_0) = 2\pi\delta(k_0 - k)$ we have

$$\widetilde{\gamma}(k) = 2\pi\widetilde{u}_s(k - k_0) - \widetilde{u}_s^*(k) \odot \widetilde{u}_s(k), \quad \mid k \mid << k_0$$

or (ignoring scaing by 2π)

$$\widetilde{u}_s(k) =$$

$$\widetilde{p}(k)\widetilde{\gamma}(k + k_0) + \widetilde{p}(k)[\widetilde{u}_s^*(k + k_0) \odot \widetilde{u}_s(k + k_0)], \quad \forall k$$

where $\widetilde{p}(k)$ is some lowpass filter with a bandwidth significantly less than k_0. Thus, upon takling the inverse Fourier transform, we obtain an expression for the (demodulated) scattered field u_s given by

$$u_s(x,k_0) = p(x) \otimes [\gamma(x)\exp(-ik_0x)]$$
$$+p(x) \otimes [\mid u_s(x,k_0) \mid^2 \exp(-ik_0x)]$$

where $p(x)$ is the inverse Fourier transform of $\widetilde{p}(k)$. Here, $p(x)$ described the bandlimited pulse that is incident on a layered dielectric described by $\gamma(x)$. The first term is a description for the weak scattered field and the second term describes the effects of multiple scattering. In terms of the 'standard model' for a stationary signal $s(t)$ as a function of time t where

$$s(t) = p(t) \otimes f(t) + n(t) \tag{21}$$

it is now clear that the Impulse Response Function (IRF) f is given by

$$f(t) = \gamma(t)\exp(-i\omega_0 t)$$

and the noise $n(t)$ is given by

$$n(t) = p(t) \otimes [\mid s(t) \mid^2 \exp(-i\omega_0 t)]$$

where ω_0 is the angular (carrier) frequency. Note that in practice, $n(t)$ will include additional background noise and in this sense, we have extract a component of the noise term that is attributed to multiple scattering effects, albeit under the conditions that: (i) $\mid u_i + u_s \mid^{-2} \sim 1$; (ii) $\mid k \mid << k_0$.

6.6 Applications to Synthetic Aperture Radar

The principles of Synthetic Aperture Radar (SAR) is discussed in Chapter 4. SAR is a side-band pulse-echo system which utilizes the response of a scatterer as it passes through the radar beam to synthesize the lateral (azimuth) resolution. This allows relatively high resolution images to be obtained at a long range. The antenna emits a pulse of microwave radiation toward the ground and the return signal is recorded at fixed time intervals along the flight path.

By studying the response of a SAR to a point scatterer in range, and then in azimuth, the point spread function of the system is established which is given by

$$p(x,y) = LT\text{sinc}(\alpha Tx)\text{sinc}(\beta k_0 y)$$

and is identical to the diffraction pattern produced by a rectangular aperture. Thus, the (post-processed) SAR image data $\hat{f}(x,y)$ generated by scattering from the ground is given by the convolution of the object function for the ground $f(x,y)$ with the appropriate point spread function, i.e.

$$\hat{f}(x,y) = p(x,y) \otimes_2 f(x,y)$$

where $\otimes_2$ denotes the two-dimensional convolution integral and $\hat{f}$ is taken to be complex data generated by f. A SAR image is usually generated by displaying the amplitude modulations of the data, i.e.

$$I_{SAR}(x,y) = \mid \hat{f}(x,y) \mid .$$

The object function describes the imaged properties of the ground surface. A conventional model for this function is the point scattering model where the object function is taken to be a distribution of point scatterers each of which reflects a replica of the emitted pulse and responds identically in azimuth. Here, nothing is said about the true physical nature of the ground surface such as its shape and material (dielectric) properties. Moreover, this standard convolution model for the data is based on application of the Born approximation where multiple scattering effects are taken to be part of the noise function $n(x,y)$. SAR images are coherent images (i.e. based on complex data containing magnitude and phase information) and consequently, contain noise that is characteristic of speckle patterns (coherent noise).

Based on results presented in Section 6.4, we consider a model for the 'ground truth' estimate $\hat{f}(x,y)$ given by

$$\hat{f}(x,y) = p(x,y) \otimes_2 [f(x,y) + \mid s(x,y) \mid^2 \exp(-ik_0 x)]$$

where the second term (on the right hand side) is taken to be the multiple scattering term[1]. Figure 23 shows the effect of applying a Gaussian lowpass filter to the complex data $\hat{f}$ before generation of the image $\mid \hat{f}(x,y) \mid$ using data obtained from the Sandia National Laboratories SAR database [30]. Filtering the complex data is undertaken in order to attempt to suppress the term $p(x,y) \otimes_2 [\mid s(x,y) \mid^2 \exp(-ik_0 x)]$. Applying a lowpass filter prior to generating an amplitude image of the complex data illiminates the cross terms generated by computing the image $\mid \hat{f}(x,y) \mid$, thereby reducing speckle, a conventional approach to speckle reduction being to apply a filter to the amplitude image. In this sense, the multiple scattering model developed in this chapter confirms a well known principle with regard to coherent image engineering which is that it is better to process complex data directly rather than the amplitude data indirectly.

6.7 Discussion

The extension of the standard model $s(t) = p(t) \otimes f(t) + n(t)$ to include multiple scattering effects that are compounded in a single term such that

$$s(t) = p(t) \otimes f(t) + p(t) \otimes [\mid s(t) \mid^2 \exp(-i\omega_0 t)] + n'(t)$$

provides a method of processing signals that is compatible with a standard signal processing 'toolkit'. Here the term $n'(t)$ include all forms of noise that does not include multiple scattering effects.

[1] After application of conventional SAR signal processing.

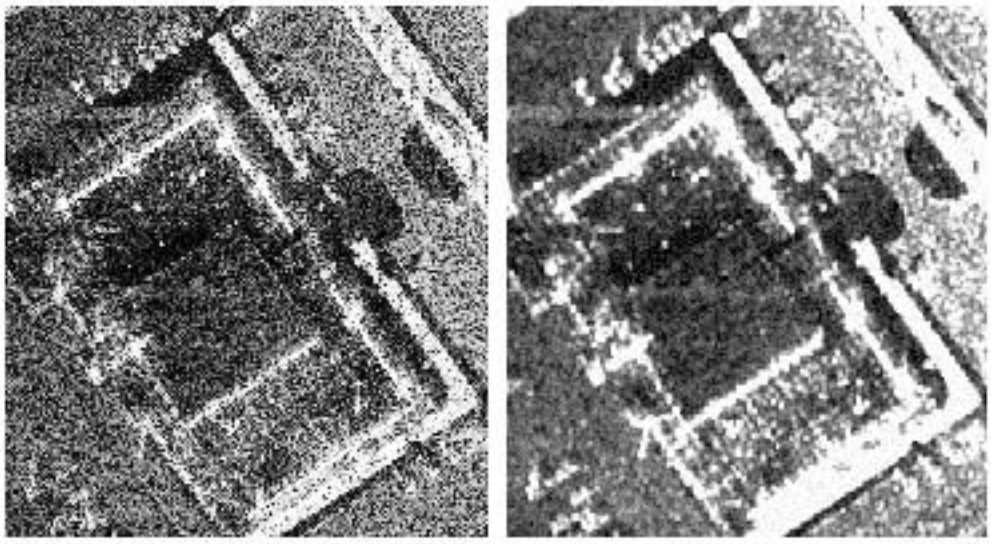

FIGURE 23 Example of an original SAR image (left) and the same image after applying a lowpass filter to the complex data (right). In both cases, the images have been histogram equalized utilizing the MATLAB function *histeq*.

The inverse scattering problem is usually formulated by first solving the forward scattering problem as discussed in Chapter 3. Once the relationship between the scattered field and the scattering function has been established by solving the Helmholtz equation, an inverse solution is then attempted. The approach taken here has been to work directly with the Helmholtz equation to produce an expression for the scattering function (as derived in Appendix 1). The example numerical simulations illustrated in Figures 19 - 22, provides evidence of the expected superiority of this solution over the weak scattering solution. However, it should be noted that this result is based on the application of the narrow side-band condition $\mid k \mid << k_0$ and that $\mid u_i(x,k) + u_s(x,k) \mid^2 \sim 1$. In this sense, the solutions developed are not strictly based on an 'exact' inverse scattering solution, but rather a modified version of the exact solution derived in Appendix 1 tailored for applicability to narrow side-band systems. Such systems are consistent with the applications of electromagnetic signal processing which is the focus of this chapter.

The approach taken in this paper has been to consider a theoretically 'exact solution' to the inverse scattering problem (as compounded in Appendix 1) and then modify this result to accommodate conditions that reduce the solution to a form that is practically realisable in terms of existing signal processing models. This is a different approach to that which is traditionally taken where a forward scattering solution is developed based on physical conditions to obtain a (forward scattering) transform which can then be inverted. With regard to pulse-echo side-band systems, our approach provides a model for strong scattering that is compounded in a single additional term. The form of this term indicates the use of lowpass filtering applied to the complex data of a SAR rather than to the image itself which is the more usual practice with regard to speckle reduction.

In general, the approach reported in this paper may provide a framework for

developing 'strong' scattering solutions that are of practical value to signal processing systems. For example, a further development of the approach used can be undertaken by relaxing the condition $u = \exp(i\theta_u)$ and considering an expansion of the term

$$\frac{1}{\mid u_i + u_s \mid^2} = 1 - u_i u_s^* - u_i^* u_s - \mid u_s \mid^2 + \ldots$$

Finally, this paper is primarily based on a one-dimensional analysis of the problem. However, the same method can be applied in two- and three-dimensions (at least for the Schrödinger and Helmholtz equations) and for completeness, the underlying results are given in Appendix 1.

7 SCATTERING FROM RANDOM MEDIA AND CLASSICAL DIFFUSION MODELS

The use of formal scattering methods for modelling the interaction of light with an inhomogeneous medium together with associated inverse scattering models is well known (e.g. [31]). In applications associated with the processing and analysis of an electromagnetic image, the aim is to develop a model that maps the object plane to the image plane. If the scattering is 'weak' (i.e. based on single scattering events) and the scattered wavefield is measured in the far field, then the map is determined by the Fourier transform which yields the fundamental imaging equation [31]

$$I(x,y) = p(x,y) \otimes_2 f(x,y) + n(x,y)$$

for an image I where p is the Point Spread Function (a characteristic of the imaging system), f is the object function and $\otimes_2$ denotes the two-dimensional convolution operation, i.e.

$$p(x,y) \otimes_2 f(x,y) = \int\int p(x-x',y-y')f(x',y')dx'dy'$$

The noise n is taken to be a stochastic function which at best, can be characterized by a probability density function $\mathrm{Pr}[n(x,y)]$ that conforms to a physically significant statistical model. The function n is taken to include a range of perturbations to the scattered field that is recorded in the image plane. Within the context of the weak scattering approximation used to derive the fundamental imaging equation, this includes multiple scattering.

The object function $f(x,y)$ is related to a three-dimensional scattering function $\gamma(\mathbf{r})$ where $\mathbf{r}$ is the three-dimensional spatial vector. In the far field, the weak scattered wavefield u_s is (ignoring scaling factors) given by the Fourier transform of the scattering function

$$u_s(\mathbf{k}) \sim \mathcal{F}_3[\gamma(\mathbf{r})]$$

where $\mathcal{F}_3$ denotes the three-dimensional Fourier transform operator and $\mathbf{k}$ is the spatial frequency vector. The inverse scattering problem is then compounded in the inversion of this result, i.e. the inverse Fourier transform. This weak scattering result

can be interpreted in terms of single scattering events generated by a scattering function consisting of an ensemble of localized point-like scatterers, for example. When multiple is present, this simple result is not sufficient to model the scattered field which must be modified to take into account double, triple, quadruple etc. scattering events. This yields results that make the objective of 'engineering' a practically viable imaging and image processing model for various applications rather intractable. In such cases, it can be of value to develop a stochastic model for the scattered field whereby, instead of relating the scattering function to some object function which is then mapped onto the image plane, we attempt to generate a model for the probability density function of a multiple scattered wavefield in order to account for the statistical distribution of the intensity field obtained in the image plane. This involves an approach in which the resultant scattered wavefield (i.e. the wave amplitude) is taken to be a consequence of a random walk where each node in the random walk is taken to be a scattering event.

7.1 Random Born Scattering

Analysis of scattering from a random medium ideally requires a model for the physical behaviour of the random variable(s) that is derived from basic principles. Ideally, this involves modelling the scattered field in terms of its interaction with an ensemble of 'scattering sites' based on an assumed stochastic process. If the density of these scattering sites is low enough so that multiple scattering is minimal, then we can apply Born scattering to develop a model for the intensity of a wavefield interacting with a random Born scatterer.

In the far field, the Born scattered field (i.e. the scattering amplitude) is given by the Fourier transform of the scattering function. If this function is known *a priori*, then the scattering amplitude can be determined. This is an example of a deterministic model. If the scattering function is stochastic (i.e. a randomly distributed scatterer) such that it can only be quantified in terms of a statistical distribution (i.e. the probability density function (PDF) - denoted by Pr) then we can simulate the (Born) scattered field by designing a random number generator that outputs deviates that conform to this distribution. The Fourier transform of this stochastic field then provides the Born scattering amplitude. Thus, given a three dimensional Helmholtz scattering function $\gamma(\mathbf{r}),\ \ \mathbf{r} \in V$ with $\Pr[\gamma(\mathbf{r})]$ known *a priori*, the scattering amplitude A is given by

$$A(\hat{\mathbf{N}}, k) = k^2 \int_V \exp(-ik\hat{\mathbf{N}} \cdot \mathbf{r})\gamma(\mathbf{r}) d^3\mathbf{r}$$

where $\hat{\mathbf{N}} = \hat{\mathbf{n}}_s - \hat{\mathbf{n}}_i$ and $\gamma(\mathbf{r})$ is a stochastic function whose deviates conform to the PDF $\Pr[\gamma(\mathbf{r})]$.

7.1.1 Random Scatterer Model

If we consider the object function f (i.e. a two-dimensional map of the three-dimensional scattering function) to be a stochastic function, then we can model this function in terms of a random distribution of amplitudes using a random number generator. A coherent image of this function is then given by (e.g. [32], [33])

$$I(x,y) =\mid p(x,y) \otimes_2 f(x,y) \mid^2$$

and an incoherent image by

$$I(x,y) =\mid p(x,y) \mid^2 \otimes_2 \mid f(x,y) \mid^2$$

where p is the Point Spread Function (PSF) for a coherent image and $\mid p \mid^2$ is the intensity PSF for an incoherent image. An example of simulating such images is given in Figure 24 which is based on the application of a zero mean Gaussian distributed random field for the object function f and Point Spread Functions for a square aperture. There is a striking difference between these images. The coherent image yields 'speckle' which is a feature of all coherent images and is due to the 'phase mixing' of the functions p and f associated with the convolution operation given above.

7.1.2 Power Spectrum Modelling

The intensity of a random Born scattered field is given by

$$I(\hat{\mathbf{N}},k) =\mid A(\hat{\mathbf{N}},k) \mid^2 = A(\hat{\mathbf{N}},k)A^*(\hat{\mathbf{N}},k)$$

$$= k^4 \int_V \exp(-ik\hat{\mathbf{N}}\cdot\mathbf{r})\gamma(\mathbf{r})d^3\mathbf{r} \int_V \exp(ik\hat{\mathbf{N}}\cdot\mathbf{r}')\gamma^*(\mathbf{r}')d^3\mathbf{r}'.$$

Using the autocorrelation theorem, we have

$$I(\hat{\mathbf{N}},k) = k^4 \int_V \exp(-ik\hat{\mathbf{N}}\cdot\mathbf{r})\Gamma(\mathbf{r})d^3\mathbf{r}$$

where Γ is the autocorrelation function given by

$$\Gamma(\mathbf{r}) = \int_V \gamma(\mathbf{r}')\gamma^*(\mathbf{r}'+\mathbf{r})d^3\mathbf{r}'.$$

This result allows us to evaluate the intensity of the Born scattered amplitude by computing the Fourier transform of the autocorrelation function of the scattering function which is taken to be composed of a number of scatterers distributed at random throughout V. This requires the autocorrelation function to be defined for a particular type of random scatterer. Thus, a random medium can be characterized via its autocorrelation function by measuring the scattered intensity and inverse Fourier transforming the result.

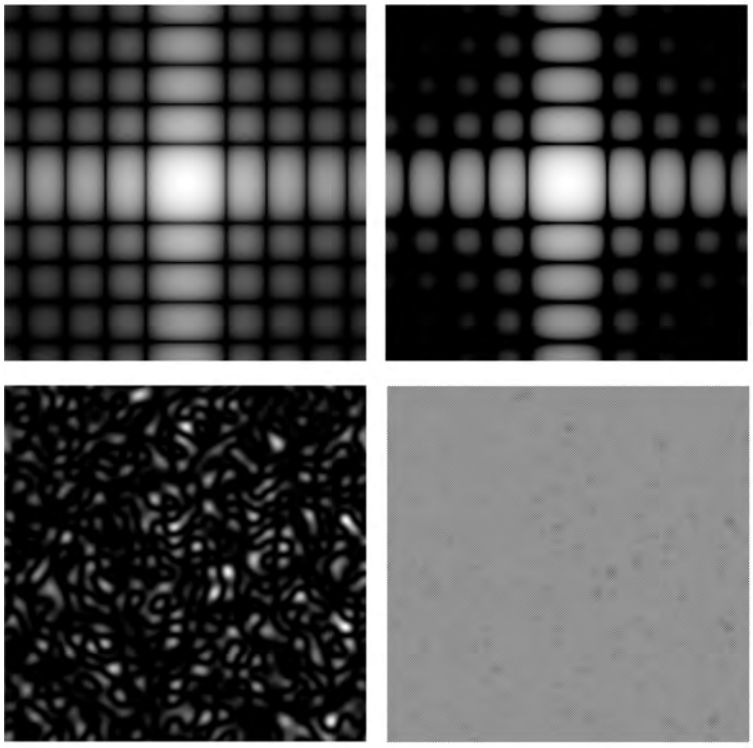

FIGURE 24 Simulation of the coherent (bottom-left) and incoherent (bottom-right) images associated with light scattering from a random medium imaged through a square aperture with coherent (top-left) and incoherent (top-right) Point Spread Functions whose absolute values are shown using a logarithmic grey-scale.

From the autocorrelation theorem, the characteristics of the autocorrelation function can be formulated by considering its expected spectral properties since

$$\Gamma(\mathbf{r}) \Longleftrightarrow \mid \widetilde{\gamma}(\mathbf{k}) \mid^2$$

where $\widetilde{\gamma}$ is the Fourier transform of γ, $\mathbf{k}$ is the spatial frequency vector and $\Longleftrightarrow$ denotes the transformation from real space $\mathbf{r}$ to Fourier space $\mathbf{k}$. Hence, in order to evaluate the most likely form of the autocorrelation function we can consider the properties of the power spectrum of the scattering function. If this function is 'white' noise, for example (i.e. its Power Spectral Density Function or PSDF is a constant), then the autocorrelation function is a delta function whose Fourier transform is a constant. However, in practice, we can expect that few scattering functions have a PSDF characterized by white noise, rather, the PSDF will tend to decay as the frequency increases. We can consider a model for the PSDF based on the Gaussian function

$$\mid \widetilde{\gamma}(\mathbf{k}) \mid^2 = \widetilde{\gamma}_0^2 \exp\left(-\frac{k^2}{k_0^2}\right),$$

for example, where $\widetilde{\gamma}_0 = \widetilde{\gamma}(0)$, $k = \mid \mathbf{k} \mid$ and k_0 is the standard deviation which is a measure of the correlation length. This form yields an autocorrelation function which is of the same type, i.e. a Gaussian function. If the geometry of the scattering function is self-affine, then we can model the scattering function as a random scattering fractal whose PSDF is characterized by [34]

$$\mid \widetilde{\gamma}(\mathbf{k}) \mid^2 \sim \frac{1}{k^{2q}}$$

where $q > 0$, the autocorrelation function being characterized by

$$\Gamma(\mathbf{r}) \sim \frac{1}{r^{3-q}}.$$

Other issues in determining the nature of the autocorrelation function are related to the physical conditions imposed on the stochastic characteristics of the scatterer.

The method discussed above can be used to model the (Born) scattered intensity from a random medium which requires an estimate of the autocorrelation of the scattering function to be known. However, this approach assumes that the density of scattering sites from which the scatterer is composed is low so that the Born approximation is valid. When the density of scattering sites increases and multiple scattering is present, the problem become progressively intractable. One approach to overcoming this problem is to resort to a purely stochastic approach which involves developing a statistical model, not for the scattering function, but for the scattered field itself which is discussed in the following section.

7.2 Statistical Modelling of the Scattered Field

Random walk methods are used as the basis for generating stochastic scattering models where the scattering of a wavefield from one scattering site to another is

taken to be a random walk in the (complex) plane with arbitrary amplitude and phase variations. We consider the wavefield E (e.g. the electric field) to be given by

$$E = \sum_{j=1}^{N} r_j \exp(i\phi_j) = R\exp(i\Phi)$$

where r, ϕ and N are independent random variables. Both r and ϕ are assumed to be continuous random variables and N is discrete. We can write E as a vector, whose components are the real and imaginary parts of E, i.e.

$$\mathbf{E} = (E_{\text{real}}, E_{\text{imag}}).$$

It is useful to work in terms of the characteristic function of a complex random variable

$$\mathbf{U} = (U_{\text{real}}, U_{\text{imag}})$$

defined as (2D inverse Fourier transform)

$$C(\mathbf{U}) = \langle \exp(i\mathbf{E}\cdot\mathbf{U})\rangle = \int \exp(i\mathbf{E}\cdot\mathbf{U})P(\mathbf{E})d\mathbf{E}$$

where the integral is taken to over all $\mathbf{E}$ and where $P(\mathbf{E})$ is the Probability Density Function (PDF) of $\mathbf{E}$. Thus, P can be computed from C via the 2D Fourier transform, i.e.

$$P(\mathbf{E}) = \frac{1}{(2\pi)^2}\int \exp(-i\mathbf{E}\cdot\mathbf{U})C(\mathbf{U})d\mathbf{U}$$

where the integral is taken over all $\mathbf{U}$.

The aim of this calculation is to find an expression for P. This is done by first computing $C(\mathbf{U}) = \langle \exp(i\mathbf{E}\cdot\mathbf{U})\rangle$ and then taking the inverse Fourier transform to evaluate P. The calculation of the characteristic function will be based on the following assumptions: (i) The phase is uniformly distributed which represents strong scattering; (ii) the scattering events at each site are independent; (iii) N conforms to a negative binomial distribution of the form

$$P_N = \begin{pmatrix} N+\alpha-1 \\ N \end{pmatrix} \frac{(\bar{N}/\alpha)^N}{(1+\bar{N}/\alpha)^{N+\alpha}}$$

where $\bar{N}$ is the mean of the distribution and α is a 'bunching' parameter. Clearly $\alpha > \bar{N}$ for P_N to be a proper PDF. Assumption (iii) above is based on a birth-death-migration processes which is representative of the distribution of scatterers.

To find $\langle \exp(i\mathbf{E}\cdot\mathbf{U})\rangle$ we write $\mathbf{E}$ and $\mathbf{U}$ in terms of their real and imaginary components, i.e.

$$\mathbf{E} = (R\cos\Phi, R\sin\Phi), \quad \mathbf{U} = (U\cos\chi, U\sin\chi)$$

where $U \equiv \mid \mathbf{U} \mid$. Here R is the resultant amplitude and Φ is the resultant phase that is detected:

$$\mathbf{E}\cdot\mathbf{U} = R\cos\Phi U\cos\chi - R\sin\Phi U\sin\chi$$

$$= U\sum_{j=1}^{N} r_j(\cos\phi_j\cos\chi - \sin\phi_j\sin\chi) = U\sum_{j=1}^{N} r_j\cos(\phi_j+\chi).$$

Hence, the characteristic function for a random walk with N steps is

$$C_N(\mathbf{U}) = \langle\exp[iU\sum_{j=1}^{N} r_j\cos(\phi_j+\chi)]\rangle.$$

Since

$$\exp(x_1+x_2+...+x_N) = \exp(x_1)\exp(x_2)\exp(x_3)...\exp(x_N),$$

$$C_N(\mathbf{U}) = \langle\prod_{j=1}^{N}\exp[iUr_j\cos(\phi_j+\chi)]\rangle.$$

The variables r, ϕ and N are independent. Assumption (ii) given above means that r_j is independent of r_k, i.e. a scattering event at site j is independent of a scattering event at site k. The net effect of this assumption is to eliminate conditional probabilities from the scattering process. In this case, the product can be taken outside the average, giving

$$C_N(\mathbf{U}) = \prod_{j=1}^{N}\langle\exp[iUr_j\cos(\phi_j+\chi)]\rangle.$$

The term $\langle\exp[iUr_j\cos(\phi_j+\chi)\rangle$ is an average over both the amplitude distribution and the phase distribution. Assuming that the phases are uniformly distributed (strong scattering), the integral for the phase can be written as

$$\langle\exp[iUr_j\cos(\phi_j+\chi)]\rangle_\phi = \int_{\forall\phi}\exp(iUr_j\cos(\phi+\chi)P_j(\phi)d\phi$$

where P_j is the uniform phase distribution defined as

$$P_j(\phi) = \begin{cases}\frac{1}{2\pi}, & -\pi\le\phi<\pi;\\ 0, & \text{otherwise.}\end{cases}$$

Consider the integral

$$I = \int_{-\pi}^{\pi}\exp[iUr_j\cos(\phi+\chi)]d\phi.$$

To evaluate this integral we use the following identity

$$\exp(i\alpha\cos\theta) = J_0(\alpha) + 2\sum_{k=1}^{\infty} i^k J_k(\alpha)\cos k\theta$$

where J_k is the Bessel function of order k. Then

$$I = \int_{-\pi}^{\pi}\left[J_0(\alpha) + 2\left(\sum_{k=1}^{\infty} i^k J_k(\alpha)\cos k\theta\right)\right]d\theta$$

$$= [J_0(\alpha)\theta]_{-\pi}^{\pi} + \left[2\sum_{k=1}^{\infty}\frac{i^k}{k}J_k(\alpha)\sin k\theta\right]_{-\pi}^{\pi} = 2\pi J_0(\alpha).$$

Hence,

$$\langle \exp(i\mathbf{E}\cdot\mathbf{U})\rangle_\phi = \langle \exp[iUr_j\cos(\phi_j+\chi)]\rangle_\phi = J_0(Ur_j)$$

where

$$U = \sqrt{U_{\text{real}}^2 + U_{\text{imag}}^2}$$

and

$$C_N(\mathbf{U}) = \prod_{j=1}^{N}\langle J_0(Ur_j)\rangle_r$$

where

$$\langle J_0(Ur_j)\rangle_r = \int_0^\infty J_0(Ur)P_j(r)dr.$$

Here $P_j(r)$ is the PDF for r. Now, if all the scattering processes are similar, then they will all have the same PDF and therefore

$$\prod_{j=1}^{N}\langle J_0(Ur_j)\rangle_r = \langle J_0(Ur)\rangle_r^N = \left(\int_0^\infty J_0(Ur)P(r)dr\right)^N.$$

This result depends on the number of steps N which is itself a random variable, and, in order to proceed further, we must consider a PDF for N. For this purpose we consider the negative binomial distribution - assumption (iii) - and develop an expression for the characteristic function for the mean $\bar{N}$ of N. This is given by

$$C_{\bar{N}}(\mathbf{U}) = \sum_{N=0}^{\infty}P_N C_N(\mathbf{U})$$

$$= \sum_{N=0}^{\infty}\begin{pmatrix} N+\alpha-1 \\ N \end{pmatrix}\frac{(\bar{N}/\alpha)^N}{(1+\bar{N}/\alpha)^{N+\alpha}}\langle J_0(Ur)\rangle_r^N$$

$$= \sum_{N=0}^{\infty}\frac{(N+\alpha-1)!}{N!(\alpha-1)!}\left(\frac{(\bar{N}/\alpha)\langle J_0(Ur)\rangle_r}{1+\bar{N}/\alpha}\right)^N\frac{1}{(1+\bar{N}/\alpha)^\alpha}$$

$$= \frac{1}{(\alpha-1)!(1+\bar{N}/\alpha)^\alpha}\sum_{N=0}^{\infty}\frac{(N+\alpha-1)!}{N!}\mu^N$$

where

$$\mu = \frac{(\bar{N}/\alpha)\langle J_0(Ur)\rangle_r}{1+\bar{N}/\alpha}.$$

Now,

$$\sum_{N=0}^{\infty}\frac{(N+\alpha-1)!}{N!}\mu^N$$

$$= (\alpha-1)!\left(1+\alpha\mu+\frac{\alpha(1+\alpha)}{2!}\mu^2+\ldots\right) = (\alpha-1)!(1-\mu)^{-\alpha}$$

and therefore we can write

$$C_{\bar{N}}(\mathbf{U}) = \frac{(\alpha-1)!}{(\alpha-1)!(1+\bar{N}/\alpha)^{\alpha}} \frac{1}{(1-\mu)^{\alpha}}$$

$$= \frac{(1+\bar{N}/\alpha)^{\alpha}}{(1+\bar{N}/\alpha)^{\alpha}(1+\bar{N}/\alpha-(\bar{N}/\alpha)\langle J_0(Ur)\rangle_r)^{\alpha}}$$

$$= \left(1+\frac{\bar{N}}{\alpha}\left(1-\langle J_0(Ur)\rangle_r\right)\right)^{-\alpha}.$$

The calculation of $\langle J_0(Ur)\rangle_r$ is based on a small but important modification whereby we scale r according to $r \to r/\sqrt{\bar{N}}$. Thus, we consider

$$\langle J_0(Ur)\rangle_r = \int_0^{\infty} P(r) J_0(Ur/\sqrt{\bar{N}}) dr.$$

As $\bar{N} \to \infty$, this modification of the definition of $\langle J_0(Ur)\rangle_r$ allows us to employ the Frobenius series for J_0, i.e.

$$J_0(x) = 1 - \frac{x^2}{4} + \frac{x^4}{2^6} - ...$$

then

$$\langle J_0(Ur)\rangle_r$$

$$= \int_0^{\infty} P(r)dr - \frac{1}{4}\int_0^{\infty} \frac{U^2r^2}{\bar{N}} P(r)dr + \frac{1}{2^6}\int_0^{\infty} \frac{U^4r^4}{\bar{N}^2} P(r)dr - ...$$

$$= 1 - \frac{1}{4}\frac{U^2}{\bar{N}}\langle r^2\rangle + \frac{1}{2^6}\frac{U^4}{\bar{N}^2}\langle r^4\rangle - ...$$

where

$$\langle r^n\rangle = \int_0^{\infty} r^n P(r) dr.$$

Hence, we can write

$$C_{\bar{N}}(\mathbf{U})$$

$$= \left[1+\frac{\bar{N}}{\alpha}\left(1-\left(1-\frac{1}{4}\frac{U^2}{\bar{N}}\langle r^2\rangle + \frac{1}{2^6}\frac{U^4}{\bar{N}^2}\langle r^4\rangle - ...\right)\right)\right]$$

$$= \left(1+\frac{1}{4}\frac{U^2}{\alpha}\langle r^2\rangle - \frac{1}{2^6}\frac{U^4}{\bar{N}\alpha}\langle r^4\rangle + ...\right)^{-\alpha}$$

and

$$C(\mathbf{U}) = \lim_{\bar{N}\to\infty} C_{\bar{N}}(\mathbf{U}) = \left(1+\frac{1}{4}\frac{U^2}{\alpha}\langle r^2\rangle\right)^{-\alpha}.$$

This result allows us to compute the PDF of $\mathbf{E} = R\exp(i\Phi)$ which can be obtained by evaluating the Fourier integral of $C(\mathbf{U})$, i.e.

$$P(\mathbf{E}) = \frac{1}{(2\pi)^2}\int_{\forall \mathbf{U}} \exp(-i\mathbf{E}\cdot\mathbf{U}) C(\mathbf{U}) d\mathbf{U}$$

$$= \frac{1}{(2\pi)^2} \int_{-\pi}^{\pi} \int_{0}^{\infty} \frac{\exp(-i\mathbf{E}\cdot\mathbf{U})}{\left(1+\frac{1}{4}\frac{U^2}{\alpha}\langle r^2\rangle\right)^{\alpha}} U dU d\chi.$$

Integrating over χ generates a Bessel function as before

$$P(\mathbf{E}) = \frac{1}{2\pi} \int_{0}^{\infty} \frac{U J_0(UR)}{\left(1+\frac{1}{4}\frac{U^2}{\alpha}\langle r^2\rangle\right)^{\alpha}} dU.$$

Evaluating the final integral gives

$$P(\mathbf{E}) = \frac{1}{2\pi 2^{\alpha-1}} \frac{R^{\alpha-1}}{\Gamma(\alpha)} \left(\frac{4\alpha}{\langle r^2\rangle}\right)^{\frac{1+\alpha}{2}} K_{\alpha-1}\left[R\left(\frac{4\alpha}{\langle r^2\rangle}\right)^{\frac{1}{2}}\right]$$

where $K_{\alpha-1}$ is a modified Bessel function. The PDF of the amplitude follows by integrating $P(\mathbf{E})$ over all values of the phase Φ. However, $P(\mathbf{E})$ is independent of Φ and so this integral yield 2π, i.e.

$$P(R) = \int_{-\pi}^{\pi} P(\mathbf{E}) R d\Phi = 2\pi R P(\mathbf{E})$$

$P(R)$ can therefore be written as

$$P(R) = \frac{\beta^{1+\alpha}}{2^{\alpha-1}\Gamma(\alpha)} R^{\alpha} K_{\alpha-1}(\beta R)$$

where

$$\beta = \left(\frac{4\alpha}{\langle r^2\rangle}\right)^{\frac{1}{2}}.$$

This is the so called 'K-distribution' whose calculation illustrates the way in which the PDF of an image can be derived subject to a model for the distribution of the phase (in this case, a uniform phase distribution representing strong scattering) and a statement of the characteristics of the random walk (in this case, a negative binomial distribution for the number of steps N). The PDF derived can then be used to characterize a signal or image (that has been generated by strong and coherent scattering processes) statistically by computing the parameters α, β and $\langle r^2\rangle$. Although this approach may be of value to the statistical analysis of a signal/image, it does not provide a solution to the inverse scattering problem. For this purpose, diffusion models for strong scattering are required as discussed below.

7.3 Derivation of the Diffusion Equation from the Wave Equation

Consider the three-dimensional homogeneous time dependent wave equation

$$\nabla^2 u - \frac{1}{c^2}\frac{\partial^2}{\partial t^2} u = 0$$

where c is taken to be a constant (light speed). Let

$$u(x,y,z,t) = \phi(x,y,z,t)\exp(i\omega t)$$

where it is assumed that field ϕ varies significantly slowly in time compared with $\exp(i\omega t)$ and note that

$$u^*(x,y,z,t) = \phi^*(x,y,z,t)\exp(-i\omega t)$$

is also a solution to the wave equation. Differentiating

$$\nabla^2 u = \exp(i\omega t)\nabla^2\phi,$$

and

$$\frac{\partial^2}{\partial t^2}u = \exp(i\omega t)\left(\frac{\partial^2}{\partial t^2}\phi + 2i\omega\frac{\partial\phi}{\partial t} - \omega^2\phi\right)$$
$$\simeq \exp(i\omega t)\left(2i\omega\frac{\partial\phi}{\partial t} - \omega^2\phi\right)$$

when

$$\left|\frac{\partial^2\phi}{\partial t^2}\right| << 2\omega\left|\frac{\partial\phi}{\partial t}\right|.$$

Under this condition, the wave equation reduces to

$$(\nabla^2 + k^2)\phi = \frac{2ik}{c}\frac{\partial\phi}{\partial t}$$

where $k = \omega/c$. However, since u^* is also a solution,

$$(\nabla^2 + k^2)\phi^* = -\frac{2ik}{c}\frac{\partial\phi^*}{\partial t}$$

and thus,

$$\phi^*\nabla^2\phi - \phi\nabla^2\phi^* = \frac{2ik}{c}\left(\phi^*\frac{\partial\phi}{\partial t} + \phi\frac{\partial\phi^*}{\partial t}\right)$$

which can be written in the form

$$\nabla^2 I - 2\nabla\cdot(\phi\nabla\phi^*) = \frac{2ik}{c}\frac{\partial I}{\partial t}$$

where $I = \phi\phi^* =\mid \phi \mid^2$. Let ϕ be given by

$$\phi(\mathbf{r},t) = A(\mathbf{r},t)\exp(ik\hat{\mathbf{n}}\cdot\mathbf{r})$$

where $\hat{\mathbf{n}}$ is a unit vector and A is the amplitude function. Differentiating, and noting that $I = A^2$, we obtain

$$\hat{\mathbf{n}}\cdot\nabla A = \frac{2}{c}\frac{\partial A}{\partial t}$$

or

$$\left(\frac{\partial}{\partial x} + \frac{\partial}{\partial y} + \frac{\partial}{\partial z}\right)A(x,y,z,t) = \frac{2}{c}\frac{\partial}{\partial t}A(x,y,z,t)$$

which is the unconditional continuity equation for the amplitude A of a wavefield

$$u(\mathbf{r},t) = A(\mathbf{r},t)\exp[i(k\hat{\mathbf{n}}\cdot\mathbf{r}+\omega t)]$$

where A varies slowly with time.

The equation

$$\nabla^2 I - 2\nabla\cdot(\phi\nabla\phi^*) = \frac{2ik}{c}\frac{\partial I}{\partial t}$$

is valid for $k = k_0 - i\kappa$ (i.e. $\omega = \omega_0 - i\kappa c$) and so, by equating the real and imaginary parts, we have

$$D\nabla^2 I + 2\mathrm{Re}[\nabla\cdot(\phi\nabla\phi^*)] = \frac{\partial I}{\partial t}$$

and

$$\mathrm{Im}[\nabla\cdot(\phi\nabla\phi^*)] = -\frac{k_0}{c}\frac{\partial I}{\partial t}$$

respectively where $D = c/2\kappa$, so that under the condition

$$\mathrm{Re}[\nabla\cdot(\phi\nabla\phi^*)] = 0$$

we obtain

$$D\nabla^2 I = \frac{\partial I}{\partial t}.$$

This is the diffusion equation for the intensity of light I. The condition required to obtain this result can be justified by applying a boundary condition on the surface S of a volume V over which the equation is taken to conform. Using the divergence theorem

$$\mathrm{Re}\int_V \nabla\cdot(\phi\nabla\phi^*)d^3\mathbf{r} = \mathrm{Re}\oint_S \phi\nabla\phi^*\cdot\hat{\mathbf{n}}d^2\mathbf{r}$$

$$= \oint_S (\phi_r\nabla\phi_r + \phi_i\nabla\phi_i)\cdot\hat{\mathbf{n}}d^2\mathbf{r}.$$

Now, if

$$\phi_r(\mathbf{r},t)\nabla\phi_r(\mathbf{r},t) = -\phi_i(\mathbf{r},t)\nabla\phi_i(\mathbf{r},t), \quad \mathbf{r}\in S$$

then the surface integral is zero and

$$D\nabla^2 I(\mathbf{r},t) = \frac{\partial}{\partial t}I(\mathbf{r},t), \quad \mathbf{r}\in V.$$

This boundary condition can be written as

$$\frac{\nabla\phi_r}{\nabla\phi_i} = -\tan\theta$$

where θ is the phase of the field ϕ which implies that the amplitude A of ϕ is constant on the boundary (i.e. $A(\mathbf{r},t) = A_0, \quad \mathbf{r}\in S, \quad \forall t$), since

$$\frac{\nabla A_0\cos\theta(\mathbf{r},t)}{\nabla A_0\sin\theta(\mathbf{r},t)} = -\frac{A_0\sin\theta(\mathbf{r},t)\nabla\theta(\mathbf{r},t)}{A_0\cos\theta(\mathbf{r},t)\nabla\theta(\mathbf{r},t)}$$

$$= -\tan\theta(\mathbf{r}, t), \quad \mathbf{r} \in S.$$

For a general field u, the homogeneous diffusion equation [35]

$$\nabla^2 u(\mathbf{r}, t) = \sigma \frac{\partial}{\partial t} u(\mathbf{r}, t), \quad \sigma = \frac{1}{D}$$

where D is the 'Diffusivity', differs in many aspects from the homogeneous scalar wave equation

$$\nabla^2 u(\mathbf{r}, t) = \frac{1}{c_0^2} \frac{\partial^2}{\partial t^2} u(\mathbf{r}, t)$$

The most important single feature is the asymmetry of the diffusion equation with respect to time. For the wave equation, if $u(\mathbf{r}, t)$ is a solution, so is $u(\mathbf{r}, -t)$. However, if $u(\mathbf{r}, t)$ is a solution of

$$\nabla^2 u = \sigma \frac{\partial u}{\partial t}$$

the function $u(\mathbf{r}, -t)$ is not; it is a solution of the quite different equation,

$$\nabla^2 u(\mathbf{r}, -t) = -\sigma \frac{\partial}{\partial t} u(\mathbf{r}, -t).$$

Thus, unlike the wave equation, the diffusion equation differentiates between past and future. This is because the diffusing field u represents the behaviour of some average property of an ensemble (e.g. of particles) which cannot in general go back to an original state. Causality must therefore be considered in the solution to the diffusion equation. This in turn leads to the use of the one-sided Laplace transform (i.e. a causal transform) for solving the equation with respect to t (compared to the Fourier transform - a non-causal transform - used to solve the wave equation with respect to t).

7.4 Green's Function Solution to the Diffusion Equation

We now consider the Green's function solution to the diffusion equation based on the Green's function derived in Chapter 3. Working in three dimensions, let us consider the general solution to the equation

$$\left(\nabla^2 - \sigma \frac{\partial}{\partial t} \right) u(\mathbf{r}, t) = -f(\mathbf{r}, t)$$

where f is a source function of compact support ($\mathbf{r} \in V$) and define the Green's function as the solution to the equation

$$\left(\nabla^2 - \sigma \frac{\partial}{\partial t} \right) G(\mathbf{r} \mid \mathbf{r}_0, t \mid t_0) = -\delta^3(\mathbf{r} - \mathbf{r}_0)\delta(t - t_0)$$

It is convenient to first take the Laplace transform of these equations with respect to $\tau = t - t_0$ to obtain

$$\nabla^2 \bar{u} - \sigma[-u_0 + p\bar{u}] = -\bar{f}$$

and

$$\nabla^2\bar{G} + \sigma[-G_0 + p\bar{G}] = -\delta^3$$

where

$$\bar{u}(\mathbf{r}, p) = \int_0^\infty u(\mathbf{r}, \tau)\exp(-p\tau)d\tau,$$

$$\bar{G}(\mathbf{r} \mid \mathbf{r}_0, p) = \int_0^\infty G(\mathbf{r} \mid \mathbf{r}_0, \tau)\exp(-p\tau)d\tau,$$

$$\bar{f}(\mathbf{r}, p) = \int_0^\infty f(\mathbf{r}, \tau)\exp(-p\tau)d\tau.$$

$$u_0 \equiv u(\mathbf{r}, \tau = 0) \text{ and } G_0 \equiv G(\mathbf{r} \mid \mathbf{r}_0, \tau = 0) = 0.$$

Pre-multiplying the equation for $\bar{u}$ by $\bar{G}$ and the equation for $\bar{G}$ by $\bar{u}$, subtracting the two results and integrating over V we obtain

$$\int_V (\bar{G}\nabla^2\bar{u} - \bar{u}\nabla^2\bar{G})d^3\mathbf{r} + \sigma\int_V u_0\bar{G}d^3\mathbf{r} = -\int_V \bar{f}\bar{G}d^3\mathbf{r} + \bar{u}(\mathbf{r}_0, p).$$

Using Green's theorem and rearranging the result gives

$$\bar{u}(\mathbf{r}_0, p) = \int_V \bar{f}(\mathbf{r}, p)\bar{G}(\mathbf{r} \mid \mathbf{r}_0, p)d^3\mathbf{r} + \sigma\int_V u_0(\mathbf{r})\bar{G}(\mathbf{r} \mid \mathbf{r}, p)d^3\mathbf{r}$$

$$+ \oint_S (\bar{g}\nabla\bar{u} - \bar{u}\nabla\bar{g}) \cdot \hat{\mathbf{n}}d^2\mathbf{r}.$$

Finally, taking the inverse Laplace transform and using the convolution theorem for Laplace transforms, we can write

$$u(\mathbf{r}_0, \tau) = \int_0^\tau \int_V f(\mathbf{r}, \tau')G(\mathbf{r} \mid \mathbf{r}_0, \tau - \tau')d^3\mathbf{r}d\tau'$$

$$+\sigma\int_V u_0(\mathbf{r})G(\mathbf{r} \mid \mathbf{r}_0, \tau)d^3\mathbf{r}$$

$$+\int_0^\tau \oint_S [G(\mathbf{r} \mid \mathbf{r}_0, \tau')\nabla u(\mathbf{r}, \tau - \tau')$$

$$-u(\mathbf{r}, \tau')\nabla G(\mathbf{r} \mid \mathbf{r}_0, \tau - \tau')] \cdot \hat{\mathbf{n}}\mathbf{d^2rd}\boldsymbol{\omega}'.$$

The first two terms are convolutions of the Green's function with the source function and the initial field $u(\mathbf{r}, \tau = 0)$ respectively.

By way of a simple example, suppose we consider the source term to be zero and the volume of interest is the infinite domain, so that the surface integral is zero. Then we have

$$u(\mathbf{r}_0, \tau) = \sigma\int_V u_0(\mathbf{r})G(\mathbf{r} \mid \mathbf{r}_0, \tau)d^3\mathbf{r}.$$

In one dimension, this reduces to

$$u(x_0,\tau)=\sqrt{\frac{\sigma}{4\pi\tau}}\int_{-\infty}^{\infty}\exp\left[-\frac{\sigma(x_0-x)^2}{4\tau}\right]u_0(x)\,dx,\quad \tau>0.$$

Observe that the field u at a time $t>0$ is given by the convolution of the field at time $t=0$ with the (Gaussian) function

$$\sqrt{\frac{\sigma}{4\pi t}}\exp\left(-\frac{\sigma x^2}{4t}\right).$$

In two-dimensions, the equivalent result is

$$u(x,y,t)=\frac{\sigma}{4\pi t}\exp\left[-\left(\frac{\sigma(x^2+y^2)}{4t}\right)\right]\otimes_2 u_0(x,y) \tag{7.1}$$

7.5 Optical Diffusion

Suppose we record the intensity I of a light field in the xy-plane for a fixed value of z. Then for $z=z_0$ say,

$$I(x,y,t)\equiv I(x,y,z_0,t)$$

so that

$$\frac{\partial}{\partial t}I(x,y,t)=D\nabla^2 I(x,y,t).$$

Let this two-dimensional diffusion equation be subject to the initial condition

$$I(x,y,0)=I_0(x,y).$$

Then, at any time $t>0$, it can be assumed that light diffusion is responsible for blurring the image I_0 and that as time increases, the image becomes progressively more (Gaussian) blurred. By comparing this model with equation (7.1) it is clear that

$$I(x,y,t)=\frac{1}{4\pi Dt}\exp\left[-\left(\frac{(x^2+y^2)}{4Dt}\right)\right]\otimes_2 I_0(x,y).$$

This result can, for example, be used to model the diffusion of light through an optical diffuser. An example of such an effect is given in Figure 25 which shows a light source (the ceiling light of a steam room) imaged through air and then through steam together with a simulation of the latter case based on the convolution of the light source with a Gaussian PSF. Steam effects light by scattering it a large number of times through the complex of small water droplets from which (low temperature) steam is composed. The high degree of multiple scattering that takes place allows us to model the transmission of light through steam in terms of a 'diffusive' rather than a 'propagative' process. The initial condition I_0 denotes the initial image which is, in effect, and with regard to Figure 25, the image of the light source obtained in air. As observed in Figure 25, the details associated with the light source are blurred through the convolution of the object function I_0 with the Gaussian PSF, a function that is characteristic of diffusion processes in general.

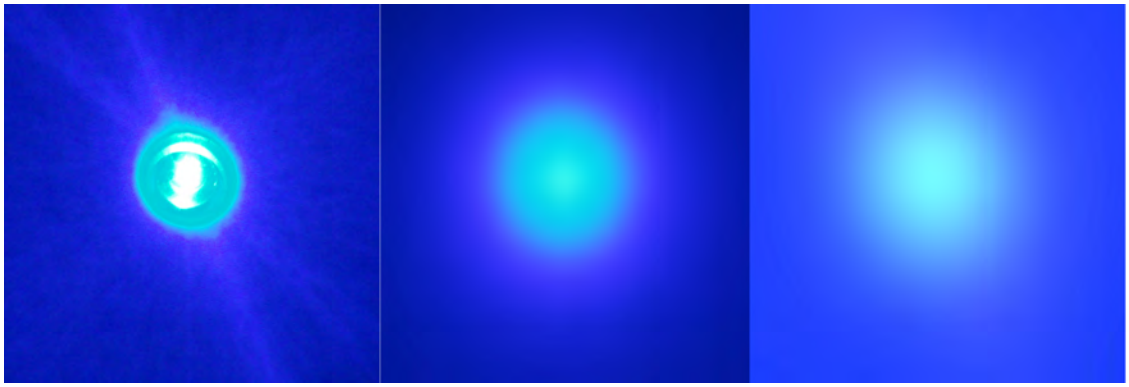

FIGURE 25 Image of an optical source (left) and the same source imaged through steam (centre) and a simulation based on the convolution of the source image with a Gaussian PSF (right).

7.6 Inverse Scattering Solutions: Dediffusion

The problem is to find I_0 from I at some time $t > 0$. Consider the case in which we record the diffused image I at a time $t = T$. The Taylor series for I at $t = 0$ may then be written as

$$I(x,y,0) = I(x,y,T) - T\left[\frac{\partial}{\partial t}I(x,y,t)\right]_{t=T}$$

$$+\frac{T^2}{2!}\left[\frac{\partial^2}{\partial t^2}I(x,y,t)\right]_{t=T} + \ldots$$

For $T << 1$, we can approximate this function be neglecting all terms after the second term. Using the diffusion equation, we then obtain

$$I(x,y,0) \simeq I(x,y,T) - T\left[\frac{\partial}{\partial t}I(x,y,t)\right]_{t=T}$$

$$= I(x,y,T) - DT\nabla^2 I(x,y,T).$$

Now, since

$$I(x,y,0) = I_0(x,y)$$

we have

$$I_0(x,y) = I(x,y,T) - DT\nabla^2 I(x,y,T).$$

7.6.1 The High Emphasis Filter

The high emphasis filter [6] is based on computing an output image I_0 from the input image I via application of the result

$$I_0(x,y) = I(x,y) - \nabla^2 I(x,y)$$

which is the case when $DT = 1$.

This filter can be implemented by computing the digital Laplacian in order to design an appropriate Finite Impulse Response (FIR) filter [33]. Applying a centre differencing scheme, i.e.

$$\nabla^2 I_{ij} = I_{(i+1)j} + I_{(i-1)j} + I_{i(j+1)} + I_{i(j-1)} - 4I_{ij}$$

we have

$$I^0_{ij} = I_{ij} - \nabla^2 I_{ij} = 5I_{ij} - I_{(i+1)j} - I_{(i-1)j} - I_{i(j+1)} - I_{i(j-1)}.$$

where

$$I^0_{ij} \equiv I_0(i,j).$$

The digital Laplacian is a shift invariant linear operation. Applying this operation to a digital image I_{ij} is the same as convolving the image with the two-dimensional array (the FIR filter)

$$\begin{pmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{pmatrix}.$$

Hence, computing I^0_{ij} is the same as convolving I_{ij} with the FIR filter

$$\begin{pmatrix} 0 & -1 & 0 \\ -1 & 5 & -1 \\ 0 & -1 & 0 \end{pmatrix}.$$

An example of the application of this filter is given in Figure 26. Given the simplicity of the process (i.e. application of a 3×3 FIR filter), the method provides an effective image enhancement technique providing the degradation of the image conforms to a light diffusion (strong scattering) model.

7.6.2 General Solution

If we record an image at a time $t = T$ then by Taylor expanding I at $t = 0$ we can write

$$I(x,y,0) = I(x,y,T) + \sum_{n=1}^{\infty} \frac{(-1)^n}{n!} T^n \left[\frac{\partial^n}{\partial t^n} I(x,y,t) \right]_{t=T}$$

The high emphasis filter derived earlier is obtained by neglecting terms in the series above for $n > 1$ giving an approximate solution for the de-diffused image I_0. If we include all the terms in this series, then an exact solution for I_0 can be obtained. This can be done by noting that (from the diffusion equation)

$$\frac{\partial^2 I}{\partial t^2} = D\nabla^2 \frac{\partial I}{\partial t} = D^2 \nabla^4 I$$

$$\frac{\partial^3 I}{\partial t^3} = D\nabla^2 \frac{\partial^2 I}{\partial t^2} = D^3 \nabla^6 I$$

and so on. In general we can write

$$\left[\frac{\partial^n}{\partial t^n} I(x,y,t) \right]_{t=T} = D^n \nabla^{2n} I(x,y,T).$$

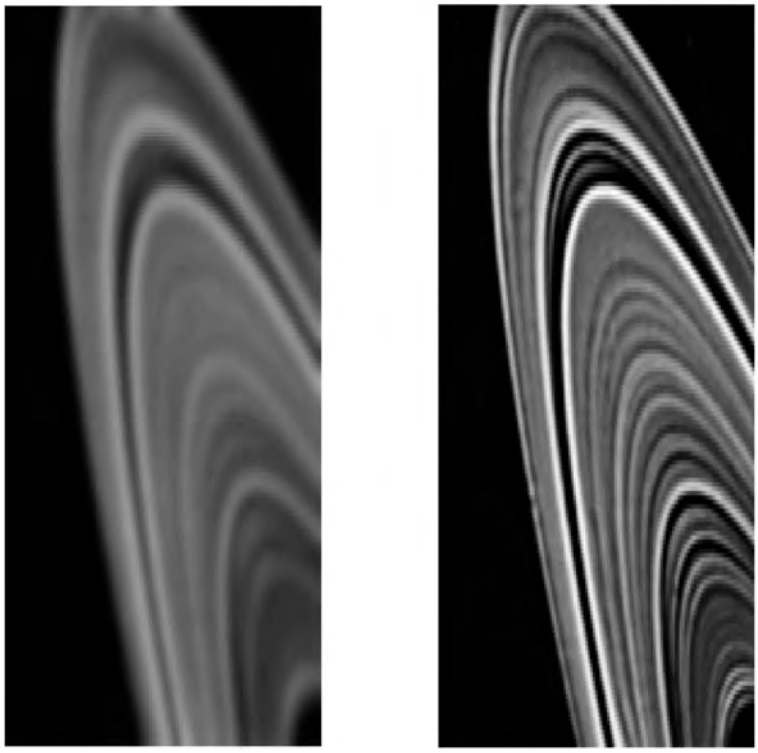

FIGURE 26 Original image (left) - rings of Saturn - and an enhanced image (right) using the high emphasis filter.

Substituting this result into the series for I_0 given above, we get

$$I_0(x,y) = I(x,y,T) + \sum_{n=1}^{\infty} \frac{(-1)^n}{n!}(DT)^n \nabla^{2n} I(x,y,T)$$

and for $DT = 1$

$$I_0 = I - \nabla^2 I + \frac{1}{2!}\nabla^4 I - \frac{1}{3!}\nabla^6 I + ...$$

From this result, we can design FIR filters for the higher order terms. Since

$$\nabla^2 I_{ij} = I_{(i+1)j} + I_{(i-1)j} + I_{i(j+1)} + I_{i(j-1)} - 4I_{ij} = J_{ij}$$

then

$$\begin{aligned}
\nabla^4 I_{ij} = \nabla^2 J_{ij} &= J_{(i+1)j} + J_{(i-1)j} + J_{i(j+1)} + J_{i(j-1)} - 4J_{ij} \\
&= I_{(i+2)j} + I_{ij} + I_{(i+1)(j+1)} + I_{(i+1)(j-1)} - 4I_{(i+1)j} \\
&\quad + I_{ij} + I_{(i-2)j} + I_{(i-1)(j+1)} + I_{(i-1)(j-1)} - 4I_{(i-1)j} \\
&\quad + I_{(i+1)(j+1)} + I_{(i-1)(j+1)} + I_{i(j+2)} + I_{ij} - 4I_{i(j+1)} \\
&\quad + I_{(i+1)(j-1)} + I_{(i-1)(j-1)} + I_{ij} + I_{i(j-2)} - 4I_{i(j-1)} \\
&\quad - 4I_{(i+1)j} - 4I_{(i-1)j} - 4I_{i(j+1)} + 4I_{i(j-1)} + 16I_{ij} \\
&= 20I_{ij} + I_{(i+2)j} + 2I_{(i+1)(j+1)} + 2I_{(i+1)(j-1)} - 8I_{(i+1)j}
\end{aligned}$$

$$+I_{(i-2)j} + 2I_{(i-1)(j+1)} + 2I_{(i-1)(j-1)} - 8I_{(i-1)j} + I_{i(j+2)}$$
$$-8I_{i(j+1)} + I_{i(j-2)} - 8I_{i(j-1)}.$$

In terms of a convolution kernel, the result above can be written as

$$\begin{pmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 2 & -8 & 2 & 0 \\ 1 & -8 & 20 & -8 & 1 \\ 0 & 2 & -8 & 2 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{pmatrix}.$$

Hence, given the convolution kernel associated with the first order solution $I - \nabla^2 I$, the convolution kernel associated with the second order solution $I - \nabla^2 I + \frac{1}{2}\nabla^4 I$ is given by

$$\begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 \\ 0 & -1 & 5 & -1 & 0 \\ 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 & \frac{1}{2} & 0 & 0 \\ 0 & 1 & -4 & 1 & 0 \\ \frac{1}{2} & -4 & 10 & -4 & \frac{1}{2} \\ 0 & 1 & -4 & 1 & 0 \\ 0 & 0 & \frac{1}{2} & 0 & 0 \end{pmatrix}$$
$$= \frac{1}{2}\begin{pmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 2 & -10 & 2 & 0 \\ 1 & -10 & 30 & -10 & 1 \\ 0 & 2 & -10 & 2 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

To compute the convolution kernel associated with the third order solution $f - \nabla^2 f + \frac{1}{2}\nabla^4 f - \frac{1}{6}\nabla^6 f$, we use the same method as above to evaluate $\nabla^6 I_{ij}$ to obtain

$$\frac{1}{6}\begin{pmatrix} 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & -3 & 15 & -3 & 0 & 0 \\ 0 & -3 & 24 & -87 & 24 & -3 & 0 \\ -1 & 15 & -87 & 202 & -87 & 15 & -1 \\ 0 & -3 & 24 & -87 & 24 & -3 & 0 \\ 0 & 0 & -3 & 15 & -3 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 & 0 \end{pmatrix}$$

An example of the application of these filters is given in Figure 27 which shows the result of diffusing a image by applying a Gaussian low-pass filter and then restoring the image using the first (high emphasis) and second order FIR filter given above.

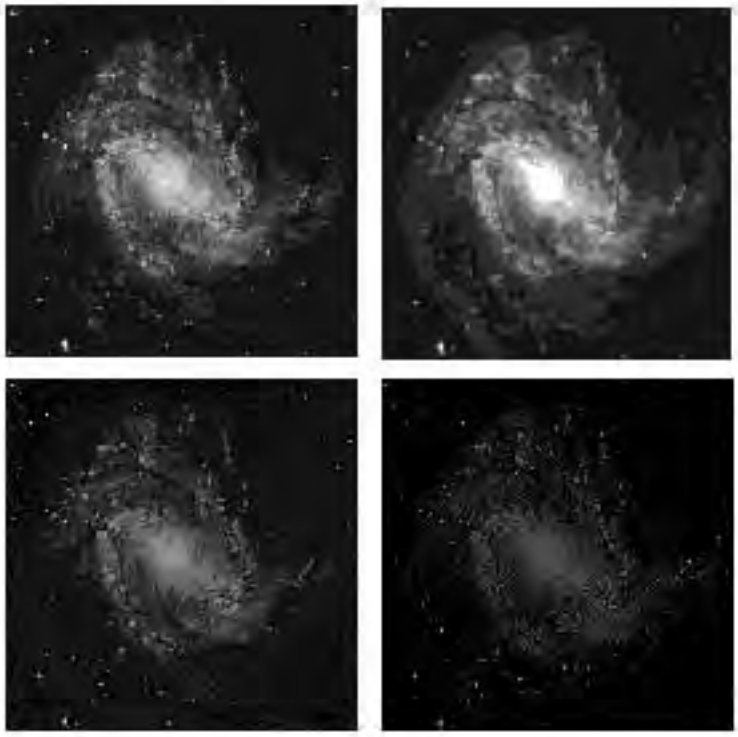

FIGURE 27 Original 256×256 image (top-left) - M83 galaxy; result after applying a Gaussian low-pass filter (top-right); output after application of the first order (high emphasis) FIR filter (bottom-left); output after application of the second order FIR filter (bottom-right).

8 FRACTIONAL DIFFUSION MODELS

8.1 Random Walk Processes

The purpose of revisiting random walk processes is that it provides a useful conceptual reference to the model that is introduced later on in this paper and in particular, appreciation of the use of the fractional diffusion equation for describing self-affine stochastic fields, an equation that arises through the unification of coherent and incoherent random walks. We shall consider a random walk in the plane where the amplitude remains constant but where the phase changes, first by a constant factor and then by a random value between 0 and 2π.

8.1.1 Coherent (Constant) Phase Walks

Consider a walk in the (real) plane where the length from one step to another is constant - the amplitude a - and where the direction that is taken after each step is the same. In this simple case, the 'walker' continues in a straight line and after n steps the total length of the path the walker has taken will be just an. We define this value as the resultant amplitude A - the total length of the walk - which will change only by account of the number of steps taken. Thus,

$$A = an.$$

If each step takes a set period of time t to complete, then it is clear that

$$A(t) = at.$$

This scenario is limited by the fact that we are assuming that each step is of precisely the same length and takes precisely the same period of time to accomplish. In general, we consider a to be the mean value of all the step lengths and t to be the cumulative time associated with the average time taken to perform all steps. A walk of this type has a coherence from one step or cluster of steps to the next, is entirely predictable and correlated in time.

If the same walk takes place in the complex plane then the phase θ from one step to the next is the same. Thus, the result is given by

$$A\exp(i\theta) = \sum_n a\exp(i\theta) = na\exp(i\theta).$$

The resultant amplitude is given by na as before and the total phase value is θ. We can also define the intensity which is given by

$$I = \mid A\exp(i\theta) \mid^2 = A^2$$

Thus, as a function of time, the intensity associated with this coherent phase walk is given by

$$I(t) = a^2t^2.$$

Suppose we make the walk slightly more complicated and consider the case where the phase increases by a small constant factor of θ at each step. After n steps, the result will be given by the sum of all the steps taken, i.e.

$$A\exp(i\Theta) = \sum_n a\exp(in\theta)$$

$$= a[1 + \exp(i\theta) + \exp(2i\theta) + ... + \exp[i(n-1)\theta]$$

$$= a\frac{[1-\exp(in\theta)]}{[1-\exp(i\theta)]}$$

$$= a\frac{\exp(in\theta/2)[\exp(-in\theta/2) - \exp(in\theta/2)]}{\exp(i\theta/2)[\exp(-i\theta/2) - \exp(i\theta/2)]}$$

$$= a\exp[i(n-1)\theta/2)]\frac{\sin(n\theta/2)}{\sin\theta/2}.$$

Now, after many steps, when n is large,

$$\alpha = (n-1)\theta/2 \simeq n\theta/2$$

and when the phase change θ is small,

$$\sin(\theta/2) \simeq \frac{\theta}{2} \simeq \frac{\alpha}{n}$$

and we obtain the result

$$A\exp(i\Theta) = na\exp[i((n-1)/2)\theta]\mathrm{sinc}\alpha, \;\; \mathrm{sinc}\alpha = \frac{\sin\alpha}{\alpha}.$$

For very small changes in the phase $\theta << 1$, $\mathrm{sinc}\alpha \simeq 1$ and the resultant amplitude A is, as before, given by an or as a function of time, by at.

8.1.2 Incoherent (Random) Phase Walk

Incoherent or random phase walks are the basis of modelling many kinds of statistical fluctuations. It is also the principle physical model associated with the stochastic behaviour of an ensemble of particles that collectively exhibit the process of diffusion. The first quantitative description of Brownian motion was undertaken by Albert Einstein and published in 1905 [36]. The basic idea is to consider a random walk in which the mean value of each step is a but where there is no correlation in the direction of the walk from one step to the next. That is, the direction taken by the walker from one step to next can be in any direction described by an angle between 0 and 360 degrees or 0 and 2π radians - for a walk in the plane. The angle that is taken at each step is entirely random and all angles are taken to be equally likely. Thus, the PDF of angles between 0 and 2π is given by

$$\Pr[\theta] = \begin{cases} \frac{1}{2\pi}, & 0 \leq \theta \leq 2\pi; \\ 0, & \text{otherwise.} \end{cases}$$

If we consider the random walk to take place in the complex plane, then after n steps, the position of the walker will be determined by a resultant amplitude A and phase angle Θ given by the sum of all the steps taken, i.e.

$$\begin{aligned} A\exp(i\Theta) &= a\exp(i\theta_1) + a\exp(i\theta_2) + ... + a\exp(i\theta_n) \\ &= a\sum_{m=1}^{n}\exp(i\theta_m). \end{aligned}$$

The problem is to obtain a scaling relationship between A and n. Clearly we should not expect A to be proportional to the number of steps n as is the case with a coherent walk. The trick to finding this relationship is to analyse the result of taking the square modulus of $A\exp(i\Theta)$. This provides an expression for the intensity I given by

$$\begin{aligned} I &= a^2\left|\sum_{m=1}^{n}\exp(i\theta_m)\right|^2 \\ &= a^2\sum_{m=1}^{n}\exp(i\theta_m)\sum_{m=1}^{n}\exp(-i\theta_m) \\ &= a^2\left[n + \sum_{j=1,j\neq k}^{n}\exp(i\theta_j)\sum_{k=1}^{n}\exp(-i\theta_k)\right]. \end{aligned}$$

Now, in a typical term

$$\exp(i\theta_j)\exp(-i\theta_k) = \cos(\theta_j - \theta_k) + i\sin(\theta_j - \theta_k)$$

of the double summation, the functions $\cos(\theta_j - \theta_k)$ and $\sin(\theta_j - \theta_k)$ have random values between ± 1. Consequently, as n becomes larger and larger, the double sum will reduces to zero since more and more of these terms cancel each other out. This insight is the basis for stating that for $n >> 1$

$$I = a^2 n$$

and the resulting amplitude is therefore given by

$$A = a\sqrt{n}.$$

In this case, A is proportional to the square root of the number of steps taken and if each step is taken over a mean time period, then we obtain the result

$$A(t) = a\sqrt{t}.$$

With a coherent walk we can state that the resulting amplitude after a time t will be at. This is a deterministic result. However, with an incoherent random walk, the interpretation of the above result is that $a\sqrt{t}$ is the amplitude associated with the most likely position that the random walker will be after time t. If we imagine many random walkers, each starting out on their 'journey' from the origin of the (complex) plane at $t = 0$, record the distances from the origin of this plane after a set period of time t, then the PDF of A will have a maximum value - the 'mode' of the distribution - that occurs at $a\sqrt{t}$. In the case of a perfectly coherent walk, the PDF will consist of a unit spike that occurs at at.

Figure 28 shows coherent and a incoherent phase walks in the plane. Each position of the walk (x_j, y_j), $j = 1, 2, 3, ..., N$ has been computed using (for $a = 1$)

$$x_j = \sum_{i=1}^{j} \cos(\theta_i)$$

$$y_j = \sum_{i=1}^{j} \sin(\theta_i)$$

where $\theta_i \in [0, 2\pi]$ is uniformly distributed and computed using the standard linear congruential pseudo random number generator

$$x_{i+1} = ax_i \text{mod} P, \quad i = 1, 2, ..., N \tag{8.1}$$

with $a = 7^7$ and $P = 2^{31} - 1$ and an arbitrary value of x_0 - the 'seed'. For the coherent phase walk

$$\theta_i = \frac{2\pi}{16} \frac{x_i}{\|\mathbf{x}\|_\infty}$$

which limits the angle to a small range between 0 and $\pi/8$ radians[1]. For the incoherent phase walk, the range of values is between 0 and 2π radians, i.e.

$$\theta_i = 2\pi \frac{x_i}{\|\mathbf{x}\|_\infty}$$

[1] $\|\mathbf{x}\|_\infty$ denote the uniform norm, equivalent to the maximum value of the array vector $\mathbf{x}$.

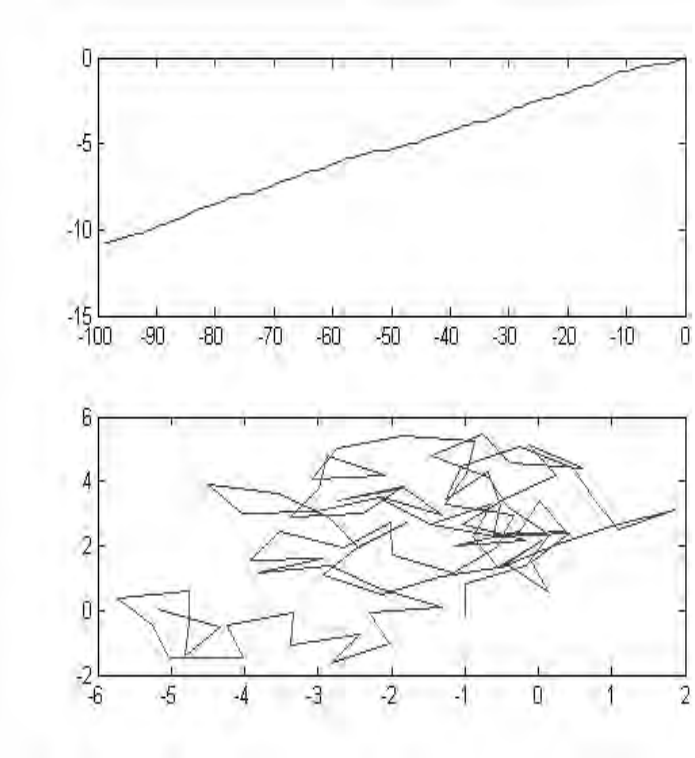

FIGURE 28 Examples of a coherent (top) and incoherent (bottom) random walk in the plane for $N = 100$.

8.2 Physical Interpretation

In the (classical) kinetic theory of matter (including gases, liquids, plasmas and some solids), we consider a to be the average distance a particle travels before it randomly collides and scatters from another particle. The scattering process is taken to be entirely elastic, i.e. the interaction does not affect the particle in any way other than to change the direction in which it travels. Thus, a represents the *mean free path* of a particle. The mean free path is a measure how far a particle can travel before scattering with another particle which in turn, is related to the number of particle per unit volume - the density of a gas for example. If we imagine a particle 'diffusing' through an ensemble of particles, then the mean free path is a measure of the 'diffusivity' of the medium in which the process of diffusion takes place. This is a feature of all classical diffusion processes which can be formulated in terms of the diffusion equation with diffusivity D. The dimensions of diffusivity are $\mathsf{length}^2/\mathsf{time}$ and may be interpreted in terms of the characteristic distance of a random walk process which varies with the square root of time.

If we consider a wavefront travelling through space and scattering from a site that changes the direction of propagation, then the mean free path can be taken to be the average number of wavelengths taken by the wavefront to propagate from one interaction to another. After scattering from many sites, the wavefront can be considered to have diffused through the 'diffuser'. Here, the mean free path

is a measure of the density of scattering sites, which in turn, is a measure of the diffusivity of the material - an optical diffuser, for example.

We can use the random walk model associated with a wavefield to interpret the flow of information through a complex network of 'sites' that are responsible for passing on the information from one site to the next. If a packet of information (e.g. a stream of bits of arbitrary length) travels directly from A to B, then, in terms of the random walk models discussed above, the model associated with this information exchange is 'propagative'; it is a coherent process which is correlated in time and its principal physical characteristic is determined by the speed at which the information flows from A to B. On the other hand, suppose that this information packet is transferred from A to B via information interchange sites C, D,...,Z,... In this case the flow of information is diffusive and is characterised by the diffusivity of the information interchange 'system'. To a first order approximation, the diffusivity will depend on the number of sites that are required to manage the reception and transmission of the information packet. As the number of sites decreases the flow of information becomes more propagative and less diffusive. Thus, we can consider the Internet, for example (albeit a good one) to be a source of information diffusion, not in terms of the diffusion of the information it coveys but in terms of the way in which information packets 'walk through' the network.

8.2.1 The Classical Diffusion Equation

The homogeneous diffusion equation is given by (for the one-dimension case x) [37]

$$\left(\frac{\partial^2}{\partial x^2} - \sigma\frac{\partial}{\partial t}\right) u(x,t) = 0$$

for a diffusivity $D = \sigma^{-1}$. The field $u(x,t)$ represents a measurable quantity whose space-time dependence is determined by the random walk of a large ensemble of particles or a multiple scattered wavefield or information flowing through a complex network. We consider an initial value for this field denoted by $u_0 \equiv u(x,0) = u(x,t)$ at $t = 0$. For example, u could be the temperature of a material that starts 'radiating' heat at time $t = 0$ from a point in space x due to a mass of thermally energised particles, each of which undertakes a random walk from the source of heat in which the most likely position of any particle after a time t is proportional to $\sqrt{t}$. In optical diffusion, for example, u denotes the intensity of light. The light wavefield is taken to be composed of an ensemble of wavefronts or rays, each of which undergoes multiple scattering as it propagates through the diffuser. For a single wavefront element, multiple scattering is equivalent to a random walk of that element.

The relationship between a random walk model and the diffusion equation can also be attributed to Einstein [36] [37] who derived the diffusion equation using a random particle model system assuming that the movements of the particles are independent of the movements of all other particles and that the motion of a single particle at some interval of time is independent of its motion at all other times. The derivation is as follows: Let τ be a small interval of time in which a particle

moves some distance between λ and $\lambda + d\lambda$ with a probability $P(\lambda)$ where τ is long enough to assume that the movements of the particle in two separate periods of τ are independent. If n is the total number of particles and we assume that $P(\lambda)$ is constant between λ and $\lambda + d\lambda$, then the number of particles which will travel a distance between λ and $\lambda + d\lambda$ in τ is given by

$$dn = nP(\lambda)d\lambda.$$

If $u(x,t)$ is the concentration (number of particles per unit volume) then the concentration at time $t+\tau$ is described by the integral of the concentration of particles which have been displaced by λ in time τ, as described by the equation above, over all possible λ, i.e.

$$u(x,t+\tau) = \int_{-\infty}^{\infty} u(x+\lambda,t)P(\lambda)d\lambda.$$

Since, τ is assumed to be small, we can approximate $u(x,t+\tau)$ using the Taylor series and write

$$u(x,t+\tau) \simeq u(x,t) + \tau\frac{\partial}{\partial t}u(x,t).$$

Similarly, using a Taylor series expansion of $u(x+\lambda,t)$, we have

$$u(x+\lambda,t) \simeq u(x,t) + \lambda\frac{\partial}{\partial x}u(x,t) + \frac{\lambda^2}{2!}\frac{\partial^2}{\partial x^2}u(x,t)$$

where the higher order terms are neglected under the assumption that if τ is small, then the distance travelled, λ, must also be small. We can then write

$$u(x,t) + \tau\frac{\partial}{\partial t}u(x,t) = u(x,t)\int_{-\infty}^{\infty} P(\lambda)d\lambda$$

$$+\frac{\partial}{\partial x}u(x,t)\int_{-\infty}^{\infty}\lambda P(\lambda)d\lambda + \frac{1}{2}\frac{\partial^2}{\partial x^2}u(x,t)\int_{-\infty}^{\infty}\lambda^2 P(\lambda)d\lambda.$$

For isotropic diffusion, $P(\lambda) = P(-\lambda)$ and so P is an even function with usual normalization condition

$$\int_{-\infty}^{\infty} P(\lambda)d\lambda = 1.$$

As λ is an odd function, the product $\lambda P(\lambda)$ is also an odd function which, if integrated over all values of λ, equates to zero. Thus we can write

$$u(x,t) + \tau\frac{\partial}{\partial t}u(x,t) = u(x,t) + \frac{1}{2}\frac{\partial^2}{\partial x^2}u(x,t)\int_{-\infty}^{\infty}\lambda^2 P(\lambda)d\lambda$$

so that

$$\frac{\partial}{\partial t}u(x,t) = \frac{\partial^2}{\partial x^2}u(x,t)\int_{-\infty}^{\infty}\frac{\lambda^2}{2\tau}P(\lambda)d\lambda.$$

Finally, defining the diffusivity as

$$D = \int_{-\infty}^{\infty} \frac{\lambda^2}{2\tau} P(\lambda) d\lambda$$

we obtain the diffusion equation

$$\frac{\partial}{\partial t} u(x,t) = D \frac{\partial^2}{\partial x^2} u(x,t).$$

8.2.2 The Classical Wave Equation

The wave equation (homogeneous form) is given by (for the one-dimension case) [38]

$$\left(\frac{\partial^2}{\partial x^2} - \frac{1}{c^2} \frac{\partial^2}{\partial t^2} \right) u(x,t) = 0$$

where c is the wave speed and u denotes the amplitude of the wavefield. A possible solution to this equation is

$$u(x,t) = p(x - ct)$$

which describes a wave with distribution p moving along x at velocity c. For the initial value problem where

$$u(x,0) = v(x), \quad \frac{\partial}{\partial t} u(x,0) = w(x)$$

the (d'Alembert) general solution is given by [38]

$$u(x,t) = \frac{1}{2}[v(x - ct) + v(x + ct)] + \frac{1}{2c} \int_{x-ct}^{x+ct} w(\xi) d\xi.$$

This solution is of limited use in that the range of x is unbounded and only applies to the case on an 'infinite string'. For the case when $w = 0$, the solution can be taken to describe two identical waves with amplitude distribution $v(x)$ travelling away from each other. Neither wave is taken to undergo any interaction as it travels along a straight path and thus, after time t the distance travelled will be ct. This is analogous to a walker undertaking a perfectly coherent walk with an average step length of c and after a period of time t reaching a position ct. The point here, is that we can relate the diffusion equation and the wave equation to two types of processes. The diffusion equation describes a field generated by incoherent random processes with no time correlations whereas the wave equation describes a field generated by coherent processes that are correlated in time. One of the aims of this paper is to formulate an equation that models the intermediate case - the fractional diffusion equation - in which random walk process have a directional bias.

8.3 Hurst Processes

For a walk in the plane, $A(t) = at$ for a coherent walk and $A(t) = a\sqrt{t}$ for an incoherent walk. However, what would be the result if the walk was neither coherent or incoherent but partial coherent/incoherent? In other words, suppose the random walk exhibited a bias with regard to the distribution of angles used to change the direction. What would be the effect on the scaling law $\sqrt{t}$? Intuitively, one expects that as the distribution of angles reduces, the corresponding walk becomes more and more coherent, exhibiting longer and longer time correlations until the process conforms to a fully coherent walk. A simulation of such an effect is given in Figure 29 which shows a random walk in the (real) plane as the (uniform) distribution of angles decreases. The walk becomes less and less random as the width of the distribution is reduced.

The equivalent effect for a random phase walk in three-dimensions is given in Figure 30. Each position of the walk

$$(x_j, y_j, z_j), \quad j = 1, 2, 3, ..., N$$

has been computed using

$$x_j = \sum_{i=1}^{j} \cos(\theta_i) \cos(\phi_i)$$

$$y_j = \sum_{i=1}^{j} \sin(\theta_i) \cos(\phi_i)$$

$$z_j = \sum_{i=1}^{j} \sin(\phi_i)$$

for $N = 500$. The uniform random number generator used to compute θ_i and ϕ_i is the same - equation (8.1) - but with different seeds. Conceptually, scaling models associated with the intermediate case(s) should be based on a generalisation of the scaling laws $\sqrt{t}$ and t to the form t^H where $0.5 \leq H < 1$. This reasoning is the basis for generalising the random walk processes considered so far, the exponent H being known as the Hurst exponent or 'dimension'.

H E Hurst (1900-1978) was an English civil engineer who designed dams and worked on the Nile river dam projects in the 1920s and 1930s. He studied the Nile so extensively that some Egyptians reportedly nicknamed him 'the father of the Nile'. The Nile river posed an interesting problem for Hurst as a hydrologist. When designing a dam, hydrologists need to estimate the necessary storage capacity of the resulting reservoir. An influx of water occurs through various natural sources (rainfall, river overflows etc.) and a regulated amount needs to be released for primarily agricultural purposes, for example, the storage capacity of a reservoir being based on the net water flow. Hydrologists usually begin by assuming that the water influx is random, a perfectly reasonable assumption when dealing with a complex ecosystem. Hurst, however, had studied the 847-year record that the Egyptians had kept of the Nile river overflows, from 622 to 1469. He noticed that large overflows tended

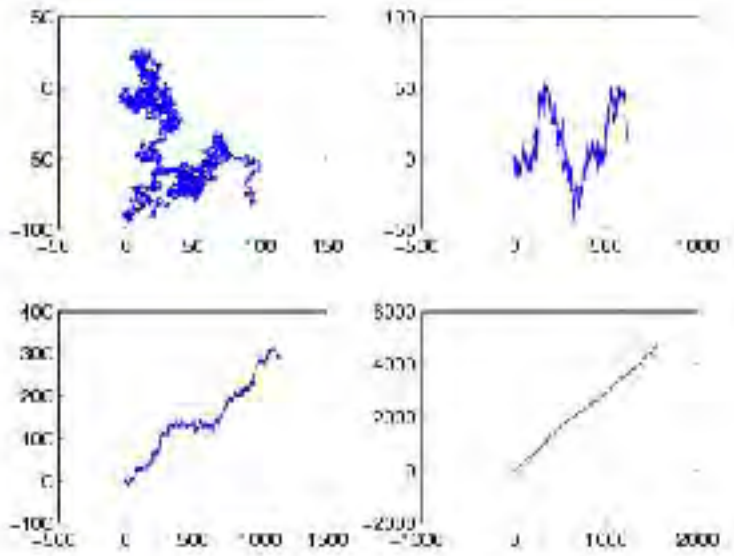

FIGURE 29 Random phase walks in the plane for a uniform distribution of angles $\theta_i \in [0, 2\pi]$ (top left), $\theta_i \in [0, 1.9\pi]$ (top right), $\theta_i \in [0, 1.8\pi]$ (bottom left) and $\theta_i \in [0, 1.2\pi]$ (bottom right).

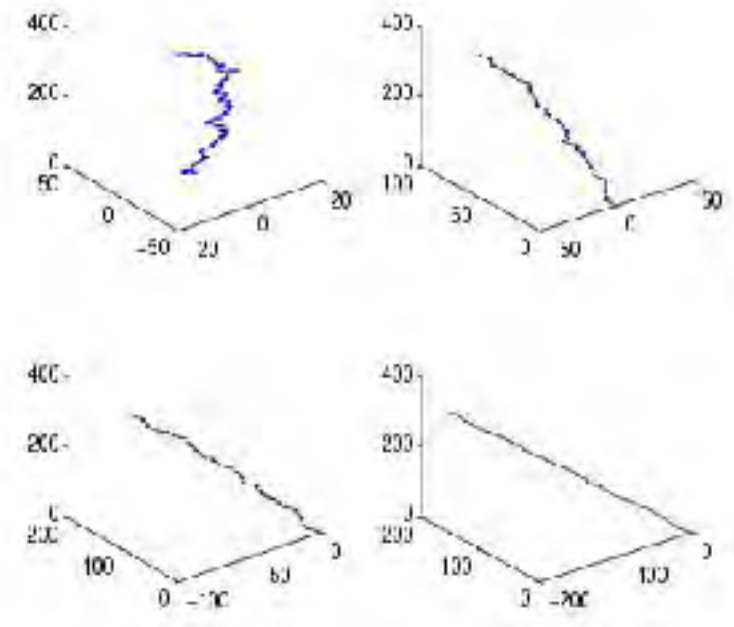

FIGURE 30 Three dimensional random phase walks for a uniform distribution of angles $(\theta_i, \phi_i) \in ([0, 2\pi], [0, 2\pi])$ (top left), $(\theta_i, \phi_i) \in ([0, 1.6\pi], [0, 1.6\pi])$ (top right), $(\theta_i, \phi_i) \in ([0, 1.3\pi], [0, 1.3\pi])$ (bottom left) and $(\theta_i, \phi_i) \in ([0, \pi], [0, \pi])$ (bottom right).

to be followed by large overflows until abruptly, the system would then change to low overflows, which also tended to be followed by low overflows. There appeared to be cycles, but with no predictable period. Standard statistical analysis of the day revealed no significant correlations between observations, so Hurst, who was aware of Einstein's work on Brownian motion, developed his own methodology [39] lead to the scaling law t^H. This scaling law makes no prior assumptions about any underlying distributions. It simply tells us how the system is scaling with respect to time. So how do we interpret the Hurst exponent? We know that $H = 0.5$ is consistent with an independently distributed system. The range $0.5 < H \leq 1$, implies a persistent time series, and a persistent time series is characterized by positive correlations. Theoretically, what happens today will ultimately have a lasting effect on the future. The range $0 < H \leq 0.5$ indicates anti-persistence which means that the time series covers less ground than a random process. In other words, there are negative correlations. For a system to cover less distance, it must reverse itself more often than a random process.

8.4 Lévy Processes

The generalisation of Einstein's equation $A(t) = a\sqrt{t}$ by Hurst to the form $A(t) = at^H, 0 < H \leq 1$ was necessary in order for Hurst to analyse the apparent random behaviour of the annual rise and fall of the Nile river for which Einstein's model was inadequate. In considering this generalisation, Hurst paved the way for an appreciation that most natural stochastic phenomena which, at first site, appear random, have certain trends that can be identified over a given period of time. In other words, many natural random patterns have a bias to them that leads to time correlations in their stochastic behaviour, a behaviour that is not an inherent characteristic of a random walk model and fully diffusive processes in general. This aspect of stochastic field theory was taken up in the late 1930s by the French mathematician Paul Lévy (1886-1971) [40].

Lévy processes are random walks whose distribution has infinite moments. The statistics of (conventional) physical systems are usually concerned with stochastic fields that have PDFs where (at least) the first two moments (the mean and variance) are well defined and finite. Lévy statistics is concerned with statistical systems where all the moments (starting with the mean) are infinite.

Many distributions exist where the mean and variance are finite but are not representative of the process, e.g. the tail of the distribution is significant, where rare but extreme events occur. These distributions include Lévy distributions. Lévy's original approach[2] to deriving such distributions is based on the following question: Under what circumstances does the distribution associated with a random walk of a few steps look the same as the distribution after many steps (except for scaling)? This question is effectively the same as asking under what circumstances do we obtain a random walk that is statistically self-affine. The characteristic function

[2] P Lévy was the research supervisor of B Mandelbrot, the 'inventor' of 'fractal geometry'.

(i.e. the Fourier transform) $P(k)$ of such a distribution $p(x)$ was first shown by Lévy to be given by (for symmetric distributions only)

$$P(k) = \exp(-a \mid k \mid^q), \quad 0 < q \leq 2$$

where a is a (positive) constant. If $q = 0$,

$$p(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp(-a) \exp(ikx) dk = \exp(-a)\delta(x)$$

and the distribution is concentrated solely at the origin as described by the delta function $\delta(x)$. When $q = 1$, the Cauchy distribution

$$p(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp(-a \mid k \mid) \exp(ikx) dk = \frac{1}{\pi} \frac{a}{a^2 + x^2}$$

is obtained and when $q = 2$, $p(x)$ is characterized by the Gaussian distribution

$$p(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp(-ak^2) \exp(ikx) dk$$

$$= \frac{1}{2\pi} \sqrt{\frac{\pi}{a}} \exp[-x^2/(4a)],$$

whose first and second moments are finite. The Cauchy distribution has a relatively long tail compared with the Gaussian distribution and a stochastic field described by a Cauchy distribution is likely to have more extreme variations when compared to a Gaussian distributed field. For values of q between 0 and 2, Lévy's characteristic function corresponds to a PDF of the form

$$p(x) \sim \frac{1}{x^{1+q}}, x \to \infty.$$

This can be shown as follows[3]: For $0 < q < 1$ and since the characteristic function is symmetric, we have

$$p(x) = \mathrm{Re}[f(x)]$$

where

$$f(x) = \frac{1}{\pi} \int_0^{\infty} e^{ikx} e^{-k^q} dk$$

$$= \frac{1}{\pi} \left(\left[\frac{1}{ix} e^{ikx} e^{-k^q} \right]_{k=0}^{\infty} - \frac{1}{ix} \int_0^{\infty} e^{ikx} (-qk^{q-1} e^{-k^q}) dk \right)$$

[3] The author acknowledges Dr K I Hopcraft, School of Mathematical Sciences, Nottingham University, England, for his help in deriving this result.

$$= \frac{q}{2\pi i x} \int_{-\infty}^{\infty} dk H(k) k^{q-1} e^{-k^q} e^{ikx}, \quad x \to \infty$$

where

$$H(k) = \begin{cases} 1, & k > 0 \\ 0, & k < 0 \end{cases}$$

For $0 < q < 1$, $f(x)$ is singular at $k = 0$ and the greatest contribution to this integral is the inverse Fourier transform of $H(k)k^{q-1}$. Noting that

$$\mathcal{F}^{-1}\left[\frac{1}{(ik)^q}\right] \sim \frac{1}{x^{1-q}}$$

where $\mathcal{F}^{-1}$ denotes the inverse Fourier transform, and that

$$H(k) \Longleftrightarrow \delta(x) + \frac{i}{\pi x} \sim \delta(x), \quad x \to \infty$$

then, using the convolution theorem, we have

$$f(x) \sim \frac{q}{i\pi x} \frac{i^{1-q}}{x^q}$$

and thus

$$p(x) \sim \frac{1}{x^{1+q}}, \quad x \to \infty$$

For $1 < q < 2$, we can integrate by parts twice to obtain

$$f(x) = \frac{q}{i\pi x} \int_0^{\infty} dk k^{q-1} e^{-k^q} e^{ikx}$$

$$= \frac{q}{i\pi x} \left[\frac{1}{ix} k^{q-1} e^{-k^q} e^{ikx}\right]_{k=0}^{\infty}$$

$$+ \frac{q}{\pi x^2} \int_0^{\infty} dk e^{ikx} [(q-1)k^{q-2} e^{-k^q} - q(k^{q-1})^2 e^{-k^q}]$$

$$= \frac{q}{\pi x^2} \int_0^{\infty} dk e^{ikx} [(q-1)k^{q-2} e^{-k^q} - q(k^{q-1})^2 e^{-k^q}], \quad x \to \infty.$$

The first term of this result is singular and therefore provides the greatest contribution and thus we can write,

$$f(x) \simeq \frac{q(q-1)}{2\pi x^2} \int_{-\infty}^{\infty} H(k) e^{ikx} (k^{q-2} e^{-k^q}) dk.$$

In this case, for $1 < q < 2$, the greatest contribution to this integral is the inverse Fourier transform of k^{q-2} and hence,

$$f(x) \sim \frac{q(q-1)}{\pi x^2} \frac{i^{2-q}}{x^{q-1}}$$

so that

$$p(x) \sim \frac{1}{x^{1+q}}, \quad x \to \infty$$

which maps onto the previous asymptotic as $q \to 1$ from the above.

For $q \geq 2$, the second moment of the Lévy distribution exists and the sums of large numbers of independent trials are Gaussian distributed. For example, if the result were a random walk with a step length distribution governed by $p(x), \ q > 2$, then the result would be normal (Gaussian) diffusion, i.e. a Brownian process. For $q < 2$ the second moment of this PDF (the mean square), diverges and the characteristic scale of the walk is lost. This type of random walk is called a Levý flight.

8.5 The Fractional Diffusion Equation

We can consider a Hurst process to be a form of fractional Brownian motion based on the generalization

$$A(t) = at^H, \quad H \in (0,1].$$

Given that incoherent random walks describe processes whose macroscopic behaviour is characterised by the diffusion equation, then, by induction, Hurst processes should be characterised by generalizing the diffusion operator

$$\frac{\partial^2}{\partial x^2} - \sigma \frac{\partial}{\partial t}$$

to the fractional form

$$\frac{\partial^2}{\partial x^2} - \sigma^q \frac{\partial^q}{\partial t^q}$$

where $q \in (0,2]$ and $D = 1/\sigma$ is the fractional diffusivity. Fractional diffusive processes can therefore be interpreted as intermediate between classical diffusive (random phase walks with $H = 0.5$; diffusive processes with $q = 1$) and 'propagative process' (coherent phase walks for $H = 1$; propagative processes with $q = 2$), e.g. [41], [42] and [43] - references therein. Fractional diffusion equations can also be used to model Lévy distributions [43] and fractal time random walks [44], [45]. However, it should be noted that the fractional diffusion operator given above is the result of a phenomenology. It is no more (and no less) than a generalisation of a well known differential operator to fractional form which follows from a physical analysis of a fully incoherent random process and it generalisation to fractional form in terms of the Hurst exponent. Note that the diffusion and wave equations can be derived rigorously from a range of fundamental physical laws (conservation of mass, the continuity equation, Fourier's law of thermal conduction, Newton's laws of motion and so on) and that, in comparison, our approach to introducing a fractional differential operator is based on postulation alone. It is therefore similar to certain other differential operators, a notable example being Schrödinger's operator.

The fractional diffusion operator given above is appropriate for modelling fractional diffusive processes that are stationary. For non-stationary fractional diffusion,

we could consider the case where the diffusivity is time variant as defined by the function $\sigma(t)$. However, a more interesting case arises when the characteristics of the diffusion processes change over time becoming less or more diffusive. This is illustrated in terms of the random walk in the plane given in Figure 31. Here, the walk starts off being fully diffusive (i.e. $H = 0.5$ and $q = 1$), changes to being fractionally diffusive ($0.5 < H < 1$ and $1 < q < 2$) and then changes back to being fully diffusive. The result given in Figure 31 shows a transition from two episodes that are fully diffusive which has been generated using uniform phase distributions whose width changes from 2π to 1.8π and back to 2π. In terms of fractional diffusion, this is equivalent to having an operator

$$\frac{\partial^2}{\partial x^2} - \sigma^q \frac{\partial^q}{\partial t^q}$$

where $q = 1, t \in (0, T_1]$; $q > 1, t \in (T_1, T_2]$; $q = 1, t \in (T_2, T_3]$ where $T_3 > T_2 > T_1$. If we want to generalise such processes over arbitrary periods of time, then we should consider q to be a function of time. We can then introduce a non-stationary fractional diffusion operator given by

$$\frac{\partial^2}{\partial x^2} - \sigma^{q(t)} \frac{\partial^{q(t)}}{\partial t^{q(t)}}.$$

This operator is the theoretical basis for non-stationary fractaional dynamic processes.

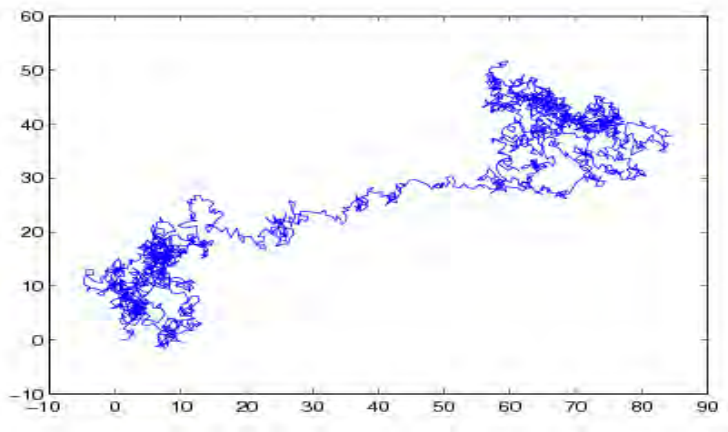

FIGURE 31 Non-stationary random phase walk in the plane.

8.6 Fractional Dynamic Model

We consider an inhomogeneous non-stationary fractional diffusion equation of the form

$$\left[\frac{\partial^2}{\partial x^2} - \sigma^{q(t)} \frac{\partial^{q(t)}}{\partial t^{q(t)}}\right] u(x,t) = F(x,t)$$

where F is a stochastic source term with some PDF and u is the stochastic field whose solution we require. Specifying q to be in the range $0 \leq q \leq 2$, leads to control over the basic physical characteristics of the equation so that we can define an anti-persistent field $u(x,t)$ when $q < 1$, a diffusive field when $q = 1$ and a propagative field when $q = 2$. In this case, non-stationarity is introduced through the use of a time varying fractional derivative whose values modify the physical characteristics of the equation.

The range of values of q is based on deriving an equation that is a generalisation of both diffusive and propagative processes using, what is fundamentally, a phenomenology. When $q = 0\ \forall t$, the time dependent behaviour is determined by the source function alone; when $q = 1\ \forall t$, u describes a diffusive process where $D = \sigma^{-1}$ is the 'diffusivity'; when $q = 2$ we have a propagative process where σ is the 'slowness' (the inverse of the wave speed). The latter process should be expected to 'propagate information' more rapidly than a diffusive process leading to transients or 'flights' of some type. We refer to q as the 'Fourier Dimension' which is related to the Hurst Exponent by $q = H + D_T/2$ where D_T is the Topological Dimension and to the Fractal Dimension D_F by $q = 1 - D_F + 3D_T/2$ as shown in Appendix 2.

Since $q(t)$ 'drives' the non-stationary behaviour of u, the way in which we model $q(t)$ is crucial. It is arguable that the changes in the statistical characteristics of u which lead to its non-stationary behaviour should also be random. Thus, suppose that we let the Fourier dimension at a time t be chosen randomly, a randomness that is determined by some PDF. In this case, the non-stationary characteristics of u will be determined by the PDF (and associated parameters) alone. Also, since q is a dimension, we can consider our model to be based on the 'statistics of dimension'. There are a variety of PDFs that can be applied which will in turn affect the range of q. By varying the exact nature of the distribution considered, we can 'drive' the non-stationary behaviour of u in different ways. However, in order to apply different statistical models for the Fourier dimension, the range of q can not be restricted to any particular range, especially in the case of a normal distribution. We therefore generalize further and consider the equation

$$\left[\frac{\partial^2}{\partial x^2} - \sigma^{q(t)}\frac{\partial^{q(t)}}{\partial t^{q(t)}}\right] u(x,t) = F(x,t), -\infty < q(t) < \infty, \forall t.$$

which allows us to apply different PDFs for q covering arbitrary ranges. For example, suppose we consider a system which is assumed to be primarily diffusive; then a 'normal' PDF of the type

$$\Pr[q(t)] = \frac{1}{\sigma\sqrt{2\pi}}\exp[-(q-1)^2/2\sigma^2], \quad -\infty < q < \infty$$

where σ is the standard deviation, will ensure that u is entirely diffusive when $\sigma \to 0$. However, as σ is increased in value, the likelihood of $q = 2$ (and $q = 0$) becomes larger. In other words, the standard deviation provides control over the likelihood of the process becoming propagative.

Irrespective of the type of distribution that is considered, the equation

$$\left[\frac{\partial^2}{\partial x^2} - \sigma^{q(t)}\frac{\partial^{q(t)}}{\partial t^{q(t)}}\right] u(x,t) = F(x,t)$$

poses a fundamental problem which is how to define and work with the term

$$\frac{\partial^{q(t)}}{\partial t^{q(t)}} u(x,t).$$

Given the result (for constant q)

$$\frac{\partial^q}{\partial t^q} u(x,t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} (i\omega)^q U(x,\omega)\exp(i\omega t)d\omega$$

we might generalize as follows:

$$\frac{\partial^{q(\tau)}}{\partial t^{q(\tau)}} u(x,t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} (i\omega)^{q(\tau)} U(x,\omega)\exp(i\omega t)d\omega.$$

However, if we consider the case where the Fourier dimension is a relatively slowly varying function of time, then we can legitimately consider $q(t)$ to be composed of a sequence of different states $q_i = q(t_i)$. This approach allows us to develop a stationary solution for a fixed q over a fixed period of time. Non-stationary behaviour can then be introduced by using the same solution for different values of q over fixed (or varying) periods of time and concatenating the solutions for all q.

8.7 Green's Function Solution

We consider a Green's function solution to the equation

$$\left(\frac{\partial^2}{\partial x^2} - \sigma^q\frac{\partial^q}{\partial t^q}\right) u(x,t) = F(x,t), \quad -\infty < q < \infty$$

when $F(x,t) = f(x)n(t)$ where $f(x)$ and $n(t)$ are both stochastic functions. Applying a separation of variables here is not strictly necessary. However, it yields a solution in which the terms affecting the temporal behaviour of $u(x,t)$ are clearly identifiable. Thus, we require a general solution to the equation

$$\left(\frac{\partial^2}{\partial x^2} - \sigma^q\frac{\partial^q}{\partial t^q}\right) u(x,t) = f(x)n(t).$$

Let

$$u(x,t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} U(x,\omega)\exp(i\omega t)d\omega$$

and

$$n(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} N(\omega) \exp(i\omega t) d\omega.$$

Then, using the result

$$\frac{\partial^q}{\partial t^q} u(x,t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} U(x,\omega)(i\omega)^q \exp(i\omega t) d\omega$$

we can transform the fractional diffusion equation to the form

$$\left(\frac{\partial^2}{\partial x^2} + \Omega_q^2 \right) U(x,\omega) = f(x) N(\omega)$$

where we shall take

$$\Omega_q = i(i\omega\sigma)^{\frac{q}{2}}$$

and ignore the case for $\Omega_q = -i(i\omega\sigma)^{\frac{q}{2}}$. Defining the Green's function g to be the solution of (see Chapter 3)

$$\left(\frac{\partial^2}{\partial x^2} + \Omega_q^2 \right) g(x \mid x_0, \omega) = \delta(x - x_0)$$

where δ is the delta function, we obtain the following solution:

$$U(x_0,\omega) = N(\omega) \int_{-\infty}^{\infty} g(x \mid x_0, \omega) f(x) dx \tag{8.2}$$

where

$$g(x \mid x_0, k) = \frac{i}{2\Omega_q} \exp(i\Omega_q \mid x - x_0 \mid)$$

under the assumption that u and $\partial u / \partial x \to 0$ as $x \to \pm\infty$. This result reduces to conventional solutions for cases when $q = 1$ (diffusion equation) and $q = 2$ (wave equation) as shall now be shown.

8.7.1 Wave Equation Solution

When $q = 2$, the Green's function defined above provides a solution for the outgoing Green's function. Thus, with $\Omega_2 = -\omega\sigma$, we have

$$U(x_0,\omega) = \frac{N(\omega)}{2i\omega\sigma} \int_{-\infty}^{\infty} \exp(-i\omega\sigma \mid x - x_0 \mid) f(x) dx.$$

Fourier inverting and using the convolution theorem for the Fourier transform, we get

$$u(x_0,t) = \frac{1}{2\sigma} \int_{-\infty}^{\infty} dx f(x) ...$$

$$\frac{1}{2\pi}\int_{-\infty}^{\infty}\frac{N(\omega)}{i\omega}\exp(-i\omega\sigma \mid x-x_0 \mid)\exp(i\omega t)d\omega$$

$$=\frac{1}{2\sigma}\int_{-\infty}^{\infty}dxf(x)\int_{-\infty}^{t}n(t-\sigma \mid x-x_0 \mid)dt$$

which describes the propagation of a wave travelling at velocity $1/\sigma$ subject to variations in space and time as defined by $f(x)$ and $n(t)$ respectively. For example, when f and n are both delta functions,

$$u(x_0,t)=\frac{1}{2\sigma}H(t-\sigma \mid x-x_0 \mid).$$

This is a d'Alembertian type solution to the wave equation where the wavefront occurs at $t=\sigma \mid x-x_0 \mid$ in the causal case.

8.7.2 Diffusion Equation Solution

When $q=1$ and $\Omega_1=i\sqrt{i\omega\sigma}$,

$$u(x_0,t)=\frac{1}{2}\int_{-\infty}^{\infty}dxf(x)...$$

$$\frac{1}{2\pi}\int_{-\infty}^{\infty}\frac{\exp(-\sqrt{i\omega\sigma} \mid x-x_0 \mid)}{\sqrt{i\omega\sigma}}N(\omega)\exp(i\omega t)d\omega.$$

For $p=i\omega$, we can write this result in terms of a Bromwich integral (i.e. an inverse Laplace transform) and using the convolution theorem for Laplace transforms with the result

$$\int_{c-i\infty}^{c+i\infty}\frac{\exp(-a\sqrt{p})}{\sqrt{p}}\exp(pt)dp=\frac{1}{\sqrt{\pi t}}\exp[-a^2/(4t)],$$

we obtain

$$u(x_0,t)=$$

$$\frac{1}{2\sqrt{\sigma}}\int_{-\infty}^{\infty}dxf(x)\int_{0}^{t}\frac{\exp[-\sigma(x_0-x)^2/(4t_0)]}{\sqrt{\pi t_0}}n(t-t_0)dt_0.$$

Now, if for example, we consider the case when n is a delta function, the result reduces to

$$u(x_0,t)=$$

$$\frac{1}{2\sqrt{\pi\sigma t}}\int_{-\infty}^{\infty}f(x)\exp[-\sigma(x_0-x)^2/(4t)]dx,\quad t>0$$

which describes classical diffusion in terms of the convolution of an initial source $f(x)$ (introduced at time $t=0$) with a Gaussian function.

8.7.3 General Series Solution

The evaluation of $u(x_0, t)$ via direct Fourier inversion for arbitrary values of q is not possible due to the irrational nature of the exponential function $\exp(i\Omega_q \mid x - x_0 \mid)$ with respect to ω. To obtain a general solution, we use the series representation of the exponential function and write

$$U(x_0, \omega) = \frac{iM_0 N(\omega)}{2\Omega_q} \left[1 + \sum_{m=1}^{\infty} \frac{(i\Omega_q)^m}{m!} \frac{M_m(x_0)}{M_0} \right] \tag{8.3}$$

where

$$M_m(x_0) = \int_{-\infty}^{\infty} f(x) \mid x - x_0 \mid^m dx.$$

We can now Fourier invert term by term to develop a series solution. Given that we consider $-\infty < q < \infty$, this requires us to consider three distinct cases.

Solution for $q = 0$

Evaluation of $u(x_0, t)$ in this case is trivial since, from equation (8.2)

$$U(x_0, \omega) = \frac{M(x_0)}{2} N(\omega) \quad \text{or} \quad u(x_0, t) = \frac{M(x_0)}{2} n(t)$$

where

$$M(x_0) = \int_{-\infty}^{\infty} \exp(- \mid x - x_0 \mid) f(x) dx.$$

Solution for $q > 0$

Fourier inverting, the first term in equation (8.3) becomes

$$\frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{iN(\omega) M_0}{2\Omega_q} \exp(i\omega t) d\omega =$$

$$\frac{M_0}{2\sigma^{\frac{q}{2}}} \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{N(\omega)}{(i\omega)^{\frac{q}{2}}} \exp(i\omega t) d\omega$$

$$= \frac{M_0}{2\sigma^{\frac{q}{2}}} \frac{1}{(2i)^q \sqrt{\pi}} \frac{\Gamma\left(\frac{1-q}{2}\right)}{\Gamma\left(\frac{q}{2}\right)} \int_{-\infty}^{\infty} \frac{n(\xi)}{(t - \xi)^{1-(q/2)}} d\xi.$$

The second term is

$$-\frac{M_1}{2} \frac{1}{2\pi} \int_{-\infty}^{\infty} N(\omega) \exp(i\omega t) d\omega = -\frac{M_1}{2} n(t).$$

The third term is

$$-\frac{iM_2}{2.2!}\frac{1}{2\pi}\int_{-\infty}^{\infty} N(\omega)i(i\omega\sigma)^{\frac{q}{2}}\exp(i\omega t)d\omega = \frac{M_2\sigma^{\frac{q}{2}}}{2.2!}\frac{d^{\frac{q}{2}}}{dt^{\frac{q}{2}}}n(t)$$

and the fourth and fifth terms become

$$\frac{M_3}{2.3!}\frac{1}{2\pi}\int_{-\infty}^{\infty} N(\omega)i^2(i\omega\sigma)^{q}\exp(i\omega t)d\omega = -\frac{M_3\sigma^{q}}{2.3!}\frac{d^{q}}{dt^{q}}n(t)$$

and

$$i\frac{M_4}{2.4!}\frac{1}{2\pi}\int_{-\infty}^{\infty} N(\omega)i^3(i\omega\sigma)^{\frac{3q}{2}}\exp(i\omega t)d\omega = \frac{M_4\sigma^{\frac{3q}{2}}}{2.4!}\frac{d^{\frac{3q}{2}}}{dt^{\frac{3q}{2}}}n(t)$$

respectively with similar results for all other terms. Thus, through induction, we can write $u(x_0,t)$ as a series of the form

$$u(x_0,t) =$$

$$\frac{M_0(x_0)}{2\sigma^{q/2}}\frac{1}{(2i)^q\sqrt{\pi}}\frac{\Gamma\left(\frac{1-q}{2}\right)}{\Gamma\left(\frac{q}{2}\right)}\int_{-\infty}^{\infty}\frac{n(\xi)}{(t-\xi)^{1-(q/2)}}d\xi$$

$$-\frac{M_1(x_0)}{2}n(t)+\frac{1}{2}\sum_{k=1}^{\infty}\frac{(-1)^{k+1}}{(k+1)!}M_{k+1}(x_0)\sigma^{kq/2}\frac{d^{kq/2}}{dt^{kq/2}}n(t).$$

Observe that the first term involves a fractional integral (the Riemann-Liouville integral), the second term is composed of the source function $n(t)$ alone (apart from scaling) and the third term is an infinite series composed of fractional differentials of increasing order $kq/2$. Also note that the first term is scaled by a factor involving $\sigma^{-q/2}$ whereas the third term is scaled by a factor that includes $\sigma^{kq/2}$.

Solution for $q < 0$

In this case, the first term becomes

$$\frac{1}{2\pi}\int_{-\infty}^{\infty}\frac{iN(\omega)M_0}{2\Omega_q}\exp(i\omega t)d\omega$$

$$=\frac{M_0}{2}\sigma^{\frac{q}{2}}\frac{1}{2\pi}\int_{-\infty}^{\infty}N(\omega)(i\omega)^{\frac{q}{2}}\exp(i\omega t)d\omega = \frac{M_0}{2}\sigma^{\frac{q}{2}}\frac{d^{\frac{q}{2}}}{dt^{\frac{q}{2}}}n(t).$$

The second term is the same is in the previous case (for $q > 0$) and the third term is

$$-\frac{iM_2}{2.2!}\frac{1}{2\pi}\int_{-\infty}^{\infty}\frac{N(\omega)i}{(i\omega\sigma)^{\frac{q}{2}}}\exp(i\omega t)d\omega$$

$$= \frac{M_2}{2.2!}\frac{1}{\sigma^{q/2}}\frac{1}{(2i)^q\sqrt{\pi}}\frac{\Gamma\left(\frac{1-q}{2}\right)}{\Gamma\left(\frac{q}{2}\right)}\int_{-\infty}^{\infty}\frac{n(\xi)}{(t-\xi)^{1-(q/2)}}d\xi.$$

Evaluating the other terms, by induction we obtain

$$u(x_0,t) = \frac{M_0(x_0)\sigma^{q/2}}{2}\frac{d^{q/2}}{dt^{q/2}}n(t) - \frac{M_1(x_0)}{2}n(t)$$

$$+\frac{1}{2}\sum_{k=1}^{\infty}\frac{(-1)^{k+1}}{(k+1)!}\frac{M_{k+1}(x_0)}{\sigma^{kq/2}}\frac{1}{(2i)^{kq}\sqrt{\pi}}\frac{\Gamma\left(\frac{1-kq}{2}\right)}{\Gamma\left(\frac{kq}{2}\right)}...$$

$$\int_{-\infty}^{\infty}\frac{n(\xi)}{(t-\xi)^{1-(kq/2)}}d\xi$$

where $q \equiv \mid q \mid$, $q < 0$. Here, the solution is composed of three terms: a fractional differential, the source term and an infinite series of fractional integrals of order $kq/2$. Thus, the roles of fractional differentiation and fractional integration are reversed as q changes from being greater than to less than zero. All fractional differential operators associated with the equations above and hence forth should be considered in terms of the definition for a fractional differential given by

$$\hat{D}^q f(t) = \frac{d^n}{dt^n}[\hat{I}^{n-q}f(t)], \quad n-q>0$$

where $\hat{I}$ is the fractional integral operator (the Riemann-Liouville transform),

$$\hat{I}^p f(t) = \frac{1}{\Gamma(p)}\int_{-\infty}^{t}\frac{f(\xi)}{(t-\xi)^{1-p}}d\xi, \quad p>0 \tag{8.4}$$

The reason for this is that direct fractional differentiation can lead to divergent integrals. However, there is a deeper interpretation of this result that has a synergy with the issue over whether a fractional diffusive system has 'memory' and is based on observing that the evaluation of a fractional differential operator depends on the history of the function in question. Thus, unlike an integer differential operator of order n, a fractional differential operator of order q has 'memory' because the value of $\hat{I}^{q-n}f(t)$ at a time t depends on the behaviour of $f(t)$ from $-\infty$ to t via the convolution with $t^{(n-q)-1}/\Gamma(n-q)$. The convolution process is of course dependent on the history of a function $f(t)$ for a given kernel and thus, in this context, we can consider a fractional derivative defined via the result above to have memory. In this sense, the operator

$$\frac{\partial^2}{\partial x^2} - \sigma^{q(t)}\frac{\partial^{q(t)}}{\partial t^{q(t)}}$$

decribes a process, compounded in a field $u(x,t)$, that has a non-stationary memory association with the temporal characteristics of the system it is attempting to model. This is not an intrinsic charcteristic of systems that are purely diffusive $q=1$ or propagative $q=2$.

8.7.4 Asymptotic Solutions for an Impulse

We consider a special case in which the source function $f(x)$ is an impulse so that

$$M_m(x_0) = \int_{-\infty}^{\infty} \delta(x) \mid x - x_0 \mid^m dx = \mid x_0 \mid^m .$$

This result immediately suggests a study of the asymptotic solution

$$u(t) = \lim_{x_0 \to 0} u(x_0, t) \tag{8.5}$$

$$= \begin{cases} \frac{1}{2\sigma^{q/2}} \frac{1}{(2i)^q \sqrt{\pi}} \frac{\Gamma\left(\frac{1-q}{2}\right)}{\Gamma\left(\frac{q}{2}\right)} \int_{-\infty}^{\infty} \frac{n(\xi)}{(t-\xi)^{1-(q/2)}} d\xi, & q > 0; \\ \frac{n(t)}{2}, & q = 0; \\ \frac{\sigma^{q/2}}{2} \frac{d^{q/2}}{dt^{q/2}} n(t), & q < 0. \end{cases}$$

The solution for the time variations of the stochastic field u for $q > 0$ are then given by a fractional integral alone and for $q < 0$ by a fractional differential alone. In particular, for $q > 0$, we see that the solution is based on the convolution integral (ignoring scaling)

$$u(t) = \frac{1}{t^{1-q/2}} \otimes n(t), \quad q > 0$$

where $\otimes$ denotes convolution and in ω-space (ignoring scaling)

$$U(\omega) = \frac{N(\omega)}{(i\omega)^{q/2}}.$$

This result is the conventional random fractal noise model for Fourier dimension q. Table **??** quantifies the results for different values of q with conventional name associations[4]. The field u has the following fundamental property for $q \in (0, 2)$:

$$\lambda^{q/2} \mathrm{Pr}[u(t)] = \mathrm{Pr}[u(\lambda t)].$$

This property describes the statistical self-affinity of u. Thus, the asymptotic solution considered here, yields a result that describes a random scaling fractal field characterized by a PSDF of the form $1/ \mid \omega \mid^q$ which is a measure of the time correlations in the signal.

Note that $q = 0$ defines the Hilbert transform of $n(t)$ whose spectral properties in the positive half space are identical to $n(t)$ because

$$\frac{1}{t} \otimes n(t) \Longleftrightarrow -i\pi \mathrm{sign}(\omega) N(\omega)$$

[4] Note that Brown noise conventionally refers to the integration of white noise but that Brownian motion is a form of pink noise because it classifies diffusive processes identified by the case when $q = 1$.

TABLE 3 Noise characteristics for different values of q. Note that the results given above ignore scaling factors.

q-value	t-space	ω-space (PSDF)	Name
$q = 0$	$\frac{1}{t} \otimes n(t)$	1	White noise
$q = 1$	$\frac{1}{\sqrt{t}} \otimes n(t)$	$\frac{1}{\lvert\omega\rvert}$	Pink noise
$q = 2$	$\int^{t} n(t)dt$	$\frac{1}{\omega^2}$	Brown noise
$q > 2$	$t^{(q/2)-1} \otimes n(t)$	$\frac{1}{\lvert\omega\rvert^q}$	Black noise

where

$$\text{sign}(\omega) = \begin{cases} 1, & \omega > 0; \\ -1, & \omega < 0. \end{cases}$$

The statistical properties of the Hilbert transform of $n(t)$ are therefore the same as $n(t)$ so that

$$\Pr[t^{-1} \otimes n(t)] = \Pr[n(t)].$$

Hence, as $q \to 0$, the statistical properties of $u(t)$ will 'reflect' those of n, i.e.

$$\Pr\left[\frac{1}{t^{1-q/2}} \otimes n(t)\right] = \Pr[n(t)], \quad q \to 0.$$

However, as $q \to 2$ we can expect the statistical properties of $u(t)$ to be such that the width of the PDF of $u(t)$ is reduced. This reflects the greater level of coherence (persistence in time) associated with the stochastic field $u(t)$ for $q \to 2$.

8.7.5 Other Asymptotic Solutions

A similar result to the asymptotic solution for $x_0 \to 0$ is obtained when the diffusivity is large, i.e.

$$\lim_{\sigma \to 0} u(x_0, t)$$

$$= \frac{M_0(x_0)}{2\sigma^{q/2}} \frac{1}{(2i)^q \sqrt{\pi}} \frac{\Gamma\left(\frac{1-q}{2}\right)}{\Gamma\left(\frac{q}{2}\right)} \int_{-\infty}^{\infty} \frac{n(\xi)}{(t-\xi)^{1-(q/2)}} d\xi$$

$$- \frac{M_1(x_0)}{2} n(t), \quad q > 0. \tag{8.6}$$

Here, the solution is the sum of fractal noise and white noise. Further, by relaxing the condition $\sigma \to 0$ we can consider the approximation

$$u(x_0,t) \simeq \frac{M_0(x_0)}{2\sigma^{q/2}} \frac{1}{(2i)^q\sqrt{\pi}} \frac{\Gamma\left(\frac{1-q}{2}\right)}{\Gamma\left(\frac{q}{2}\right)} \int_{-\infty}^{\infty} \frac{n(\xi)}{(t-\xi)^{1-(q/2)}} d\xi$$

$$-\frac{M_1(x_0)}{2} n(t) + \frac{M_2(x_0)}{2.2!}\sigma^{q/2} \frac{d^{q/2}}{dt^{q/2}} n(t), \quad q > 0, \quad \sigma << 1 \tag{8.7}$$

in which the solution is expressed in terms of the sum of fractal noise, white noise and the fractional differentiation[5] of white noise.

8.7.6 Equivalence with a Wavelet Transform

The wavelet transform is defined in terms of projections of $f(t)$ onto a family of functions that are all normalized dilations and translations of a prototype 'wavelet' function w [46], i.e.

$$\mathcal{W}[f(t)] = F_L(t) = \int_{-\infty}^{\infty} f(\tau) w_L(\tau, t) d\tau$$

where

$$w_L(\tau,t) = \frac{1}{\sqrt{L}} w\left(\frac{\tau - t}{L}\right), \quad L > 0.$$

The independent variables L and t are continuous dilation and translation parameters respectively. The wavelet transformation is essentially a convolution transform where $w_L(t)$ is the convolution kernel with dilation variable L. The introduction of this factor provides dilation and translation properties into the convolution integral that gives it the ability to analyse signals in a multi-resolution role (the convolution integral is now a function of L), i.e.

$$F_L(t) = w_L(t) \otimes f(t), \quad L > 0.$$

In this sense, the asymptotic solution (ignoring scaling)

$$u(t) = \frac{1}{t^{1-q/2}} \otimes n(t), \quad q > 0 \quad x \to 0$$

is compatible with the case of a wavelet transform where

$$w_1(t) = \frac{1}{t^{1-q/2}}$$

for the stationary case and where, for the non-stationary case,

$$w_1(t,\tau) = \frac{1}{t^{1-q(\tau)/2}}.$$

[5] As defined by equation (8.4).

8.8 Solution to the Fractional Diffusion Equation

Consider the fractional diffusion equation for the intensity $I(x,y,t)$ of light in the image plane z given by

$$\nabla^2 I(\mathbf{r},t) - \sigma^q \frac{\partial^q}{\partial t^q} I(\mathbf{r},t) = I_0(\mathbf{r})\delta(t)$$

where $\mathbf{r} = \hat{\mathbf{x}}x + \hat{\mathbf{y}}y$ and $I_0(\mathbf{r})$ is a source function with an impulse at $t = 0$. For $q = 1$, the solution to this equation in the infinite domain is (with $r =\mid \mathbf{r} \mid$ and $I(\mathbf{r}, t = 0) = 0$ as shown in Chapter 3)

$$I(\mathbf{r},t) = I_0(\mathbf{r}) \otimes_2 G(r,t)$$

where, for $t > 0$,

$$G(r,t) = \frac{1}{4\pi t} \exp\left[-\left(\frac{\sigma r^2}{4t}\right)\right]$$

which is the solution of

$$\left(\nabla^2 - \sigma\frac{\partial}{\partial t}\right) G(r,t) = -\delta^2(r)\delta(t).$$

For the fractional diffusion equation, we consider a similar (Green's function) solution but where the Green's function is given by the solution of

$$\left(\nabla^2 - \sigma^q\frac{\partial^q}{\partial t^q}\right) G(r,t) = -\delta^2(r)\delta(t).$$

Using the Fourier based operator for a fractional derivative, we can transform this equation into the form

$$(\nabla^2 + \Omega_q^2) g(\mathbf{r} \mid \mathbf{r}', \omega) = -\delta^2(\mathbf{r} - \mathbf{r}')$$

where

$$g(\mathbf{r} \mid \mathbf{r}', \omega) = \int_{-\infty}^{\infty} G(\mathbf{r} \mid \mathbf{r}', t) \exp(-i\omega t) dt,$$

$$\Omega_q^2 = -i\omega\sigma, \quad \Omega_q = \pm i(i\omega\sigma)^{q/2}.$$

Note that for $q = 2$, this equation becomes

$$(\nabla^2 + k^2) g(\mathbf{r} \mid \mathbf{r}', \omega) = \delta^2(\mathbf{r} - \mathbf{r}')$$

where $k = \pm\omega\sigma$. This equation defines the Green's function for the time independent wave operator in two-dimensions, the 'out going' Green's functions being given by

$$g(\mathbf{r} \mid \mathbf{r}', k) = \frac{1}{\sqrt{8\pi}} \exp(i\pi/4) \frac{\exp(ik \mid \mathbf{r} - \mathbf{r}' \mid)}{\sqrt{k \mid \mathbf{r} - \mathbf{r}' \mid}}$$

Generalizing this result, for $q \in (1,2)$, by writing the exponential function in its series form, with $R = \mid \mathbf{r} - \mathbf{r}' \mid$ we have, for $\Omega_q = i(i\omega\sigma)^{q/2}$,

$$G(R,t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} d\omega \exp(i\omega t)\frac{\exp(i\pi/4)}{\sqrt{8\pi}}\frac{\exp[-(i\omega\sigma)^{q/2}R]}{\sqrt{iR}(i\omega\sigma)^{q/4}}$$

$$= \frac{1}{\sqrt{8\pi R}}\frac{1}{2\pi}\int_{-\infty}^{\infty} d\omega \exp(i\omega t)\left(\frac{1}{(i\omega\sigma)^{q/4}} - (i\omega\sigma)^{q/4}R + \frac{1}{2!}(i\omega\sigma)^{3q/4}R^2 - ...\right)$$

$$= \frac{1}{\sqrt{8\pi R}}\frac{1}{\sigma^{q/4}t^{1-q/4}} - \sqrt{\frac{R}{8\pi}}\sigma^{q/4}\delta^{q/4}(t) + \frac{1}{\sqrt{8\pi}}\sum_{n=1}^{\infty}\frac{(-1)^{n+1}}{(n+1)!}R^{(2n+1)/2}\sigma^{3nq/4}\delta^{3nq/4}(t) \tag{8.8}$$

Simplification of this infinite sum can be addressed be considering suitable asymptotics, the most significant of which (for arbitrary values of R) is the case when the (fractional) diffusivity D is large. In particular, we note that as $\sigma \to 0$,

$$G(R,t) = \frac{1}{\sqrt{8\pi R}\sigma^{q/4}t^{1-(q/4)}}.$$

Thus, we can consider a solution to the two-dimensional fractional diffusion equation (for a tenuous medium when $\sigma \to 0$)

$$\left(\nabla^2 - \sigma^q\frac{\partial^q}{\partial t^q}\right) I(\mathbf{r},t) = I_0(\mathbf{r})\delta(t)$$

of the form

$$I(x,y) = \frac{1}{2\sqrt{2\pi}}\frac{1}{(DT)^{1-q/4}}\frac{1}{(x^2+y^2)^{\frac{1}{4}}} \otimes_2 I_0(x,y).$$

Comparing this solution with the solution to the two-dimensional diffusion equation, i.e.

$$I(x,y) = \frac{1}{4\pi DT}\exp\left[-\left(\frac{x^2+y^2)}{4DT}\right)\right] \otimes_2 I_0(x,y),$$

we observe that when the diffusivity is large and the diffusion time $t = T$ is small such that $DT = 1$, the difference between an image obtained by a full two-dimensional diffuser and a fractional diffuser is compounded in the difference between the convolution of the initial image with (ignoring scaling) the functions $\exp(-R^2/4)$ and $1/\sqrt{R}$, respectively. Compared with the Gaussian (at least for $DT \geq 1$), the function $R^{-1/2}$ decays more rapidly and hence will have broader spectral characteristics leading to an output that is less 'diffused' than that produced by the convolution of the input with a Gaussian. In terms of the fractional diffusion equation being used to model scattering in a tenuous medium, this is to be expected.

The Green's function used to derive this result is based on the condition $\sigma\omega R >> 1$ which is incompatible with use of $\sigma \to 0$ unless $\omega R \to \infty$ faster than $\sigma \to 0$. In otherwords, the approximation

$$\left(\frac{1}{(i\omega\sigma)^{q/4}} - (i\omega\sigma)^{q/4}R + \frac{1}{2!}(i\omega\sigma)^{3q/4}R^2 - ...\right) \sim \frac{1}{(i\omega\sigma)^{q/4}}, \quad R \mid \omega\sigma \mid^{\frac{q}{2}} << 1$$

needs to be consistent with the condition $\sigma\omega R >> 1$ upon which the series above is based. In order to do this more terms in the above series need to be included. Thus the Point Spread Function characterised by spectral response $R^{-1/2}$ is just the first order effect of based on the infinite series expression for $G(R,t)$ given by equation (8.8). However, this term is the only term in the series solution for G - equation (8.8) - that is non-zero forall $t > 0$, and in this sense, represents the principal (time-dependent) solution.

Noting that in two-dimensions, the Green's function is given by $-\ln(kR)$ for $kR \to 0$ (see Chapter 3), then the fractional Green's function is given by

$$g(R,\omega) = -\ln[(\sigma\omega)^{q/2}R], \quad \sigma \to 0$$

In this case, the image model is given by

$$I(x,y) = -\ln[(\sigma\omega)^{q/2}\sqrt{x^2+y^2}] \otimes_2 I_0(x,y)$$

Compared to the Gaussian function and like the function $R^{-1/2}$ the Point Spread Function $\ln(R)$ is singular and has broader spectral characteristics than a Gaussian Point Spread Function.

8.9 Inverse Solution

Let I_0 be represented as a Taylor series at some time $T > 0$, i.e.

$$I(\mathbf{r},0) = I(\mathbf{r},T) + T\left[\frac{\partial}{\partial t}I(\mathbf{r},t)\right]_{t=T} - \frac{T^2}{2!}\left[\frac{\partial^2}{\partial t^2}I(\mathbf{r},t)\right]_{t=T} + ...$$

Now, since

$$\frac{\partial u}{\partial t} = \frac{\partial^{1-q}}{\partial t^{1-q}}\frac{\partial^q}{\partial t^q}u$$

then from the fractional diffusion equation

$$\frac{\partial u}{\partial t} = D^q\frac{\partial^{1-q}}{\partial t^{1-q}}\nabla^2 u$$

and

$$\frac{\partial^2}{\partial t^2}u$$

$$= \frac{\partial}{\partial t}\left(\frac{\partial u}{\partial t}\right) = \frac{\partial}{\partial t}\left(D^q\frac{\partial^{1-q}}{\partial t^{1-q}}\nabla^2 u\right) = D^q\frac{\partial^{1-q}}{\partial t^{1-q}}\nabla^2\frac{\partial u}{\partial t}$$

$$= D^q\frac{\partial^{1-q}}{\partial t^{1-q}}\nabla^2\left(D^q\frac{\partial^{1-q}}{\partial t^{1-q}}\nabla^2 u\right) = D^{2q}\frac{\partial^{1-q}}{\partial t^{1-q}}\left(\frac{\partial^{1-q}}{\partial t^{1-q}}\nabla^4 u\right)$$

so that in general,

$$\frac{\partial^n u}{\partial t^n} = D^{nq}\frac{\partial^{n(1-q)}}{\partial t^{n(1-q)}}\nabla^{2n}u.$$

Because (see Appendix 3)

$$\frac{\partial^{-q}}{\partial t^{-q}} I(\mathbf{r},t) = \frac{1}{\Gamma(q)t^{1-q}} \otimes I(\mathbf{r},t)$$

we can write the Taylor series for the field at $t = 0$ in terms of the field at $t = T$ as

$$I(\mathbf{r},0) = I(\mathbf{r},T) + \frac{TD^q}{\Gamma(q)} \left[\frac{\partial}{\partial t} \left(\frac{1}{t^{1-q}} \otimes \nabla^2 I(\mathbf{r},t) \right) \right]_{t=T}$$

$$- \frac{T^2 D^{2q}}{2!\Gamma(2q)} \left[\frac{\partial^2}{\partial t^2} \left(\frac{1}{t^{1-2q}} \otimes \nabla^4 I(\mathbf{r},t) \right) \right]_{t=T}$$

$$+ \frac{T^3 D^{3q}}{3!\Gamma(3q)} \left[\frac{\partial^3}{\partial t^3} \left(\frac{1}{t^{1-3q}} \otimes \nabla^6 I(\mathbf{r},t) \right) \right]_{t=T} - \ldots$$

Note that for $T << 1$,

$$I(\mathbf{r},0) = I(\mathbf{r},T) + \frac{TD^q}{\Gamma(q)} \left[\frac{\partial}{\partial t} \left(\frac{1}{t^{1-q}} \otimes \nabla^2 I(\mathbf{r},t) \right) \right]_{t=T}$$

and under the condition that

$$\left[\frac{\partial}{\partial t} \left(\frac{1}{t^{1-q}} \otimes I(\mathbf{r},t) \right) \right]_{t=T} = I(\mathbf{r},T)$$

we can write

$$I(\mathbf{r},0) = I(\mathbf{r},T) + \frac{TD^q}{\Gamma(q)} \nabla^2 I(\mathbf{r},T).$$

8.10 Deconvolution

In the presence of additive noise $n(x,y)$, the deconvolution problem is as follows: Given that

$$I(x,y) = p(x,y) \otimes_2 I_0(x,y) + n(x,y)$$

where $\Pr[n(x,y)]$ is known (ideally), find an estimate for I_0. This is a common problem in optics (digital image processing) known as the deconvolution problem whose solution is fundamental to image restoration and reconstruction [47], [48]. In terms of the material presented in this paper, there are two Point Spread Functions (PSF) $p(x,y)$ that have been considered: For full diffusion (strong scattering)

$$p(x,y) = \frac{1}{4\pi DT} \exp\left[- \left(\frac{(x^2+y^2)}{4DT} \right) \right]$$

and for fractional diffusion (intermediate scattering in a tenuous medium with large diffusivity)

$$p(x,y) = \frac{1}{2\sqrt{2\pi}} \frac{1}{(DT)^{1-q/4}} \frac{1}{(x^2+y^2)^{\frac{1}{4}}}.$$

Based on the result provided in Appendix 4, we note that

$$\frac{1}{4\pi DT}\exp\left[-\left(\frac{(x^2+y^2)}{4DT}\right)\right] \leftrightarrow \exp[-4DT(k_x^2+k_y^2)]$$

and

$$\frac{1}{2\sqrt{2\pi}}\frac{1}{(DT)^{1-q/4}}\frac{1}{(x^2+y^2)^{\frac{1}{4}}}$$

$$\updownarrow$$

$$\frac{\sqrt{\pi}\Gamma(0.75)}{\Gamma(0.25)(DT)^{1-q/4}}\frac{1}{(k_x^2+k_y^2)^{3/4}}$$

where Γ denotes the Gamma function. In the latter case, the filter is a 'fractal filter' and thus, if I_0 is characterised by white noise, then the output I is a Mandelbrot surface with a fractal dimension of 2.5 [50], [51]. In the absence of noise, the inverse solution for I_0 can be written in the form (evaluating the Gamma functions)

$$I_0(x,y) = 1.67(DT)^{1-q/4}\nabla^{\frac{3}{2}}I(x,y),$$

a result that is based on the application of the fractional Laplacian or Riesz operator [49]

$$\nabla^q \leftrightarrow \mid \mathbf{k} \mid^q .$$

Figure 32 shows the effect of filtering an image using full diffusion and fractional diffusion for $DT = 1$. Comparison of the results shows that fractional diffusion does not blur the image to the same extent which is to be expected given the physical characteristics under which fractional diffusion processes are taken to occur, i.e. in terms of intermediate multiple scattering events in a tenuous rarefied medium.

There are a range of approaches to solving the one-dimensional and two-dimensional deconvolution problem in practice (i.e. with additive noise) leading to the classification of different 'inverse filters'. If *a priori* information on the statistics of the noise function and the object function is available, then Bayesian estimation methods are preferable in the design of filters whose performance will then depend on statistical parameters such as the standard deviation. In some cases, an estimate of $\Pr[n(x,y)]$ can be obtained by taking an image (and a number of images to obtain a statistically significant result) with zero input, i.e. with $I_0 = 0$. This provides a method of validating an idealised PDF through data fitting and, thus, determination of the statistical parameters from which a theoretical PDF is composed. In cases when experimental determinism is not practically possible, statistical models are used directly. This includes models such as the K-distribution discussed and derived in Section 7.2. However, with regard to incoherent imaging systems, the noise function tends to be Gaussian distributed - a result of the noise being a linear combination of many different independent noise source which combine to produce Gaussian noise (a consequence of the Central Limit Theorem).

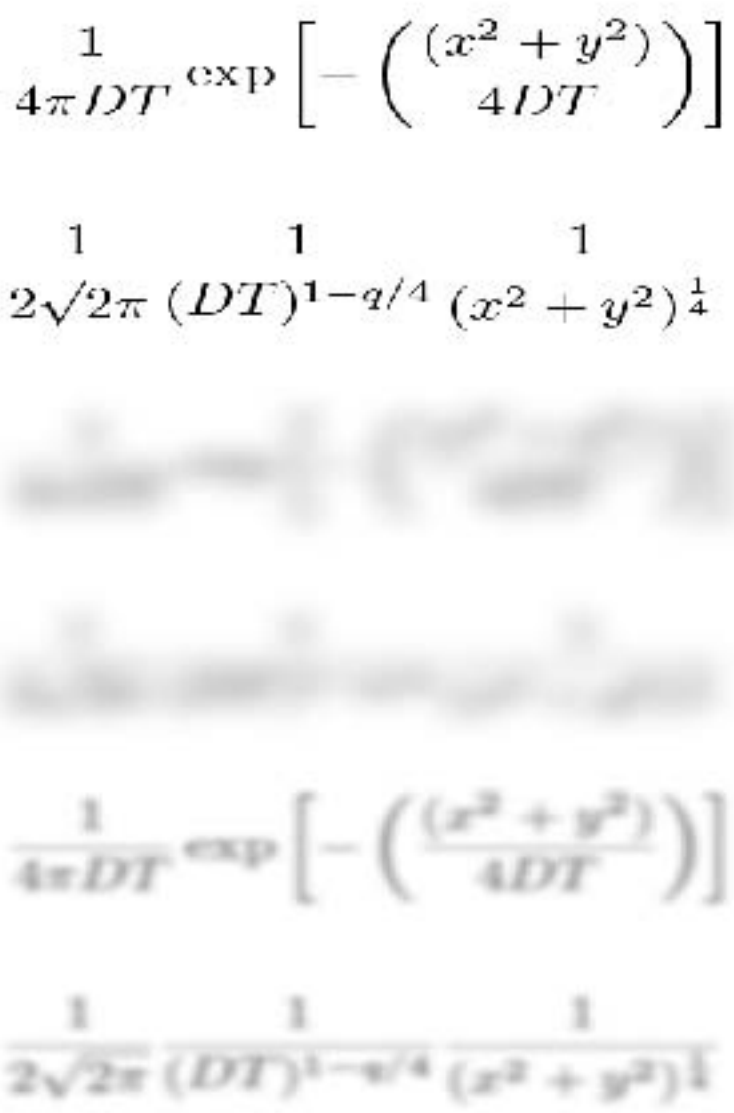

FIGURE 32 Comparison between the effect of diffusion (centre) and fractional diffusion (bottom) on a binary image (top) for $DT = 1$.

8.10.1 Bayesian Estimation

Using Bayes rule, the aim is to find an estimate for I_0 such that

$$\frac{\partial}{\partial I_0} \ln \Pr[n(x,y)] + \frac{\partial}{\partial I_0} \ln \Pr[I_0(x,y)] = 0.$$

Consider the following models for the PDFs: (i) Gaussian statistics for the noise when (ignoring scaling and where σ_n^2 is the standard deviation of n)

$$\Pr[n(x,y)] =$$

$$\exp\left(-\frac{1}{\sigma_n^2}\int\int[(I(x,y) - p(x,y) \otimes_2 I_0(x,y)]^2 dxdy\right).$$

(ii) Gaussian statistics for the object function where (ignoring scaling and where $\sigma_{I_0}^2$ is the standard deviation of I_0)

$$\Pr[I_0(x,y)] = \exp\left(-\frac{1}{\sigma_{I_0}^2}\int\int I_0^2(x,y)dxdy\right).$$

Differentiating, these statistical models yield the equation

$$I(x,y) \odot_2 p(x,y) = \frac{\sigma_n^2}{\sigma_{I_0}^2} f(x,y) + [p(x,y) \otimes_2 f(x,y)] \odot_2 p(x,y)$$

where $\odot_2$ denotes the two-dimensional correlation integral. In Fourier space, this equation becomes

$$\widetilde{I}(k_x,k_y)P^*(k_x,k_y) = \frac{1}{\Gamma^2}\widetilde{I}_0(x,y) + \mid P(k_x,k_y) \mid^2 I_0(k_x,k_y)$$

The filter $F(k_x,k_y)$ for Gaussian statistics is therefore given by

$$F(k_x,k_y) = \frac{P^*(k_x,k_y)}{\mid P(k_x,k_y) \mid^2 + \sigma_n^2/\sigma_{I_0}^2}$$

where σ_n/σ_{I_0} defines the signal-to-noise ratio of $I(x,y)$. and $\widetilde{I}_0(k_x,k_y) = F(k_x,k_y)\widetilde{I}(k_x,k_y)$. The reconstruction for I_0 is then given by

$$I_0(x,y) =$$

$$\frac{1}{(2\pi)^2}\int\int F(k_x,k_y)\widetilde{I}(k_x,k_y)\exp(ik_x x)\exp(ik_y y)dk_x dk_y$$

8.10.2 Adaptive Filtering

Given $P(k_x,k_y)$, the performance of this filter depends on the value of $\Sigma = \sigma_n^2/\sigma_{I_0}^2$. In general, as $\Sigma \to 0$ the reconstruction sharpens but at the expense of 'ringing'. Thus, an optimum value of Σ is obtained by computing I_0 over a range of values of

Σ and, for each reconstruction, computing the ratio of the number of zero crossings Z_c to the sum of the magnitude of a digital gradient $\sum \mid \mathcal{D}I_0[i,j] \mid$, i.e.

$$R = \frac{Z_c}{\sum \mid \mathcal{D}I_0[i,j] \mid}$$

This ratio is based on the principle that an optimum reconstruction is one which provides a sharp image with minimal ringing, i.e. a reconstruction for which R is a minimum. This principle has been applied in the example results given in the following section. Note that the Fourier based approach to image restoration relies on the ability to implement the convolution and correlation theorems. This requires that the data has been recorded by an (optical) imaging system that is isoplanatic (i.e. the Point Spread Function is stationary).

8.11 Example Applications: Image Enhancement in Astronomy

We consider examples of image reconstruction based on equation (8.5) for fully diffusive and fractional diffusive models using the optimization procedure discussed in the previous section for the following 'digital Laplacian'

$$\mathcal{D}I_0[i,j] = \begin{pmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{pmatrix}.$$

8.11.1 Deconvolution for Full Diffusion

Figure 33 shows the application of equation (8.5) where (ignoring scaling and with $\sigma = 4DT$)

$$P(k_x, k_y) = \exp[-\sigma(k_x^2 + k_y^2)].$$

In this example, the diffusion of the object has been generated by turbulence of the earths atmosphere through which light from the object has been fully diffused. In this case, the reconstruction depends on the value of both σ and Σ and an optimization scheme based on computing $I_0[i,j;\sigma,\Sigma]$ for $\min R$.

8.11.2 Deconvolution for Fractional Diffusion

Fractional diffusion models apply to scattering processes that occur in a tenuous and extremely rarefied medium. In applied optics, one of the most common examples of this phenomena occurs in astronomy and the processes associated with light scattering from cosmic dust which is composed of particles which are a few molecules to the order of 10^{-4} metres in size. Cosmic dust is defined in terms of its astronomical location including intergalactic dust, interstellar dust, interplanetary dust and circumplanetary dust (such as in a planetary ring). In our own Solar System, interplanetary dust is generated from sources such as comet dust, asteroidal dust, dust from the Kuiper belt and interstellar dust passing through our solar system.

FIGURE 33 Diffusion based deconvolution (below) of an image of Saturn observed by a ground based telescope with light diffused by the atmosphere (above).

This dust is responsible for zodiacal light which is produced by sunlight reflecting off dust particles. Cosmic dust can be categorised in terms of different types of nebulae associated with different physical causes and processes. These include: diffuse nebula, infrared reflection nebula, supernova remnants and molecular clouds, for example. However, in a more general sense, cosmic dust often characterises the interstellar medium which is the gas and dust that pervade interstellar space. This medium consists of an extremely dilute (by terrestrial standards) mixture of ions, atoms, molecules, and larger dust grains, consisting of about 99% gas and 1% dust by mass. Densities range from a few thousand to a few hundred million particles per cubic meter with an average value in the Milky Way Galaxy, for example, of a million particles per cubic meter. In comparison with the scattering of light from earth-based random media, for example, the interstellar medium is highly diffuse and therefore ideal for applying light scattering models based on fractional diffusion when $D \to \infty$.

Figure 34 shows the application of equation (8.6) where (ignoring scaling)

$$P(k_x, k_y) = \frac{1}{(k_x^2 + k_y^2)^{3/4}}$$

from an optical image obtained with the Hubble Space Telescope. This image is part of the constellation of Perseus as observed through an interstellar dust cloud that covers nearly 4 degrees on the sky observed 1,000 light-years away.

FIGURE 34 Fractional diffusion based deconvolution (right) for $\sigma_n/\sigma_{I_0} = 1$ of a dust clouded star field (left) in the constellation of Pegasus.

8.12 Discussion

We have considered different approaches to modelling light scattering through random media including: formal scattering methods, cross-correlation models for a scattering function under the weak field condition, statistical modelling of the wavefield and application of the diffusion equation for modelling multiple scattering processes. The formal scattering approach provides inverse solutions that are, in general, not of any practical value to signal and image processing. Cross-correlation methods are of value in modelling the intensity distribution but are not generally applicable to image processing problems. While statistical modelling methods are useful for developing theoretical PDFs of images and their statistical evaluation, they are not directly applicable for image enhancement.

The use of a fully diffusive process for modelling strong (multiple) scattering provides a result that is applicable in terms of solving the inverse scattering problem which is compounded in terms of developing a suitable deconvolution algorithm. We have extended this approach to model intermediate scattering by generalizing the diffusion equation to the fractional form

$$\left(\nabla^2 - \sigma^q \frac{\partial^q}{\partial t^q}\right) I(\mathbf{r}, t) = I_0(\mathbf{r})\delta(t)$$

where $I(x, y, t)$ is a light intensity image, D^{-1} is the fractional diffusivity and $q \in (1, 2)$. A solution has been considered based on a Fourier transform representation of a fractional derivative for which the initial condition $I(\mathbf{r}, t = 0)$ (used to solve the diffusion equation) is not required. An asymptotic result has then been derived for the case when $\sigma \to 0$ that is compounded in an Optical Transfer Function given by $(k_x^2 + k_y^2)^{-0.75}$.

9 LOW FREQUENCY ELECTROMAGNETIC SCATTERING AND A UNIFIED WAVEFIELD THEORY

We review the inhomogeneous scalar Helmholtz equation in three-dimensions and the scattering of scalar wavefields from a scatterer of compact support. An asymptotic solution is then considered representing the effect of the frequency approaching zero when a 'wavefield' reduces to a 'field'. The characteristics of ultra-low frequency Helmholtz scattering are then considered and the physical significance discussed of a model that is based on the scattering of Helmholtz wavefields over a broad frequency spectrum. This is equivalent to using a linear systems approach for modelling the propagation, interaction and detection of broad-band signals and provides an approach to the classification of a field from a wavefield that is intrinsically causal and thus, consistent with the basic principle of information theory. The approach leads to the proposal that all fields are derived from wavefields interacting over a broad frequency spectrum and that there are two principal field types: (i) fields generated by low frequency scattering - a 'gravitational field'; (ii) fields generated by high frequency eigenfield tendency - an 'electric field'.

9.1 Introduction

The ideas presented in this chapter are a first attempt to develop a universal physical model in which 'fields' and 'particles' do not exist along with such concepts as 'charge'. All that is considered is a universe consisting of scalar wavefields whose governing equation is the (inhomogeneous) Helmholtz equation over a broad frequency spectrum with a bandwidth that is determined by the Planck length

$$\ell = \sqrt{\frac{\hbar G}{c_0^3}} \sim 1.16 \times 10^{-35} \text{metres}$$

where $\hbar$ is Dirac's constant (Planck's constant divided by 2π), G is the gravitational constant and c_0 is the speed of light. The frequency associated with the Planck length is $c_0/\ell \sim 2.59 \times 10^{43}$Hz.

The rationale for a Planck bandwidth is as follows: Consider the hypothetical case where the de Broglie wavelength λ associated with a non-relativistic particle with constant velocity $v << c_0$ is continually decreased. The rest mass m of the particle will then increase according to $m = 2\pi\hbar/(v\lambda)$. As the mass increases, its Newtonian gravitational field will increase as will the escape velocity $v_e = \sqrt{2Gm/r} = \sqrt{4\pi\hbar G/(v\lambda r)}$ where r is the distance required to escape the gravitational field. Suppose that the wavelength becomes so small that the escape velocity is equal to the speed of light (i.e. the particle becomes a micro black hole), then $\lambda r = 4\pi\hbar G/(vc_0^2)$. We define the Planck length for the limiting case when $r \to 4\pi\lambda$ and $v \to c_0$, i.e. the length associated with the case when the velocity of a particle approaches the speed of light and the distance required to escape the gravitational field approaches the de Broglie wavelength of the particle. The Planck frequency sets a upper limit on the band width of a universal spectrum since, beyond this frequency, any particle (and the de Broglie wavefield associated with it) will not be detectable. The breadth of the spectrum is taken to be a consequence of the 'big-bang' (i.e. a broad frequency spectrum is the product of a short impulse).

Although the approach considered in this thesis has some philosophical similarities to string theory, which is increasingly being challenged by a number of authors (e.g. [52], [53]), it is different in its 'scale'. If string theory is concerned with the interpretation of physics through wavefields with a wavelength of the order of ℓ, here, we consider wavefields interacting (scattering) at all scales greater than the Planck length (i.e. over all frequencies less than the Planck frequency). In a sense, we consider the universe itself to be a single 'string' composed of a broad spectrum of (scalar) wavefields. This is a 'waves within waves' approach and can thus be interpreted in terms of a universal fractal model [54], not in terms of the 'shape of the universe' but in terms of the wavefields from which it is taken to be composed. Here, we adopt a formal scattering theory approach for a scalar Helmholtz wavefield and derive both standard and some non-standard results which are considered in terms of two fundamental experimental observations, the Poisson spot and the Einstein ring.

9.2 Field Equations

The field equations for electromagnetic and gravitational fields (i.e. Maxwell's equations [55] and Einstein's equations [56], respectively) appear to have only one thing in common: they both predict wave behaviour (the wavefields being composed of very different 'fields' with different properties), namely, electromagnetic waves and gravity waves respectively where, in the latter case, no direct experimental observations have been made, to date. In quantum mechanics, the quantum fields that are modelled through equations such as the Schrödinger [57], Dirac [58], [59], [60], Klein-

Gordon (e.g. [61], [62]) and Rarita-Schwinger [63] equations, are not fields in the sense of an electric (vector) field or a gravitational (tensor - a curved vector space) field but wavefields of different types, i.e. scalar (Klein-Gordon and Schrödinger equations for the relativistic and non-relativistic case, respectively), scalar-spinor (Dirac equations), vector (Proca equations [64], [65]) and vector-spinor (Rarita-Schwinger equations) fields. The theoretical origin of these wavefields is a direct result of the fundamental postulates of quantum mechanics, namely, that energy $E = \hbar\omega$ and momentum $\mathbf{p} = \hbar\mathbf{k}$ for a wavefield with (angular frequency) ω and wavenumber $\mid \mathbf{k} \mid = 2\pi/\lambda$. Relating energy and momentum (particulate concepts associated with Newtonian mechanics) to frequency and wavelength respectively immediately raises the issue of particle verses wave. It also brings into focus the question of whether a field or a wavefield is more fundamental.

Apart from the Schrödinger equation, all of the equations listed above describe relativistic quantum fields. They are all 'products' of the fact that, given the postulates of quantum mechanics, Einstein's special theory of relativity allows for the existence of scalar, scalar-spinor, vector, vector-spinor and tensor fields. In each case, the field, as characterised by a given operator, is taken to describe a 'particle' (a localised entity) that is classified in terms of a Boson or Fermion which have integer or half-integer spin (the intrinsic angular momentum) respectively. This is compounded in Table **??** (where m denotes the rest mass):

TABLE 4 Classification of different fields in terms of a Boson or Fermion

Equation name	**Field Type**	**Spin $s\hbar$**	**Example**
Klein-Gordon	Scalar	$s = 0$	Higgs boson
Dirac	Scalar Spinor	$s = 1/2$	leptons: electrons, muons
Proca-Maxwell	Vector	$s = 1$	$m = 0$: photons gluons; $m \neq 0$: mesons
Rarita-Schwinger	Vector Spinor	$s = 3/2$	None discovered
Gravitation	Tensor	$s = 2$	gravitons

Note that, like the graviton, the Higgs boson is a hypothetical particle that is taken to explain the origins of mass m which has, to date, not been verified experimentally. The terms 'Boson' and 'Fermion' relate to the fact that the statistical behaviour of integer spin particles can be classified in terms of Bose-Einstein statistics and half-integer spin particles, in terms of Fermi-Dirac statistics.

Vector bosons are considered to mediate three of the four fundamental inter-

TABLE 5 Summary of the principal physical forces, their range and example Bosons.

Force	**Range**	**Transmitted by Bosons**
Gravitational	Long	Graviton, $m = 0$, $s = 2$
Electromagnetic	Intermediate	Photon, $m = 0$, $s = 1$
Weak	Short	$W^{\pm}$, Z_0, $m \neq 0$, $s = 1$
Strong	Short	gluons, $m = 0$, $s = 1$

actions in 'particle' physics, i.e. electromagnetic, weak and strong interactions, and tensor bosons (gravitons) are assumed to mediate the gravitational force as summarised in Table **??**:

Of the four fundamental forces in nature, gravity was the first to be 'invented' but, to this day, remains the most elusive. With just criticism over his universal theory of gravity and, in particular, the principle of instantaneous action at a distance, upon which the theory is based, Isaac Newton rightly stated that '*... I have told you how it works, not why*'. Here, we consider a causal approach to explaining the 'why'.

9.3 Fields, Wavefields and the Proca Equations

In electromagnetism and general relativity, the field equations are considered to be fundamental, the wave properties of these fields being a consequence of decoupling (under certain conditions) the field equations. In other words, the wave properties of these fields are, in a sense, a by-product of writing a set of coupled equations in terms of a single or set of equations of the same (wave) type. What if a wave equation was to determine the form of the field equations and thus the characteristics of the field(s)? The first to consider such an approach was the Romanian born Alexandru Proca who derived the Proca or Proca-Maxwell equations.

For a three-dimensional space $\mathbf{r} = \hat{\mathbf{x}}x + \hat{\mathbf{y}}y + \hat{\mathbf{z}}z$, with time denoted by t and with the Laplacian operator defined as

$$\nabla^2 \equiv \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2},$$

it is well known that Maxwell's equations (specifically, the microscopic equations for point 'charges') can be decoupled to produce the inhomogeneous wave equations (e.g. [60], [66], [67])

$$\left(\nabla^2 - \frac{1}{c_0^2}\frac{\partial^2}{\partial t^2}\right)\phi(\mathbf{r}, t) = -\frac{\rho}{\epsilon_0}$$

and

$$\left(\nabla^2 - \frac{1}{c_0^2}\frac{\partial^2}{\partial t^2}\right)\mathbf{A}(\mathbf{r}, t) = -\mu_0\mathbf{j}$$

for the magnetic vector potential $\mathbf{A}$ and the electric scalar potential ϕ where ρ is the charge density, $\mathbf{j}$ is the current density and ϵ_0 and μ_0 are the permittivity and permeability of free space, respectively. This requires use of the gauge transforms

$$\mathbf{A} \rightarrow \mathbf{A} + \nabla X \text{ and } \phi \rightarrow \phi - \frac{\partial X}{\partial t}$$

where the gauge function X is taken to satisfy the homogeneous wave equation

$$\left(\nabla^2 - \frac{1}{c_0^2} \frac{\partial^2}{\partial t^2} \right) X = 0.$$

The solutions at $(\mathbf{r}_0, t_0)$ for the 'retarded potentials' ϕ and $\mathbf{A}$ are then given by

$$\phi(\mathbf{r}_0, t_0) = \frac{1}{4\pi\epsilon_0} \int \frac{\rho(\mathbf{r}, \tau)}{\mid \mathbf{r} - \mathbf{r}_0 \mid} d^3\mathbf{r}, \quad \tau = t_0 - \mid \mathbf{r} - \mathbf{r}_0 \mid / c_0$$

and

$$\mathbf{A}(\mathbf{r}_0, t_0) = \frac{\mu_0}{4\pi} \int \frac{\mathbf{j}(\mathbf{r}, \tau)}{\mid \mathbf{r} - \mathbf{r}_0 \mid} d^3\mathbf{r}$$

which show that a change in ρ and $\mathbf{j}$ affects ϕ and $\mathbf{A}$ $\mid \mathbf{r} - \mathbf{r}_0 \mid / c_0$ seconds later. The change propagates away from the sources ρ and $\mathbf{j}$ at a velocity c_0 which is the theoretical basis for the propagation of electromagnetic waves.

In quantum mechanics, energy E and momentum $\mathbf{p}$ are replaced by the wave operators

$$-i\hbar \frac{\partial}{\partial t} \text{ and } i\hbar\nabla$$

respectively. Thus, the non-relativistic 'free energy' (no potential energy component) equation

$$E = \frac{p^2}{2m}$$

yields Schrödinger's equation [68]

$$i\hbar \frac{\partial}{\partial t} U = -\frac{\hbar^2}{2m} \nabla^2 U$$

for a unit amplitude plane wave of the form

$$U(\mathbf{r}, t) = \exp[i(\mathbf{k} \cdot \mathbf{r} - \omega t)].$$

In the relativistic case when

$$E = \pm\sqrt{p^2 c_0^2 + m^2 c_0^4} \text{ or } E^2 = p^2 c_0^2 + m^2 c_0^4$$

we obtain the (homogeneous) Klein-Gordon equation [60]

$$\left(\nabla^2 - \frac{1}{c_0^2} \frac{\partial^2}{\partial t^2} \right) U - \kappa^2 U = 0$$

where $\kappa = mc_0/\hbar$. This equation is taken to describe massive scalar Bosons (spin 0 particles) such as the Higgs boson. In contrast, the classical wave equation is taken to describe massless scalar (for the electric field potential) or vector (for the magnetic vector potential) Bosons, i.e. the photon.

Given that Maxwell's equations can be decoupled to produce inhomogeneous wave equations for ϕ and $\mathbf{A}$, Proca's idea was to modify Maxwell's equations in order to produce inhomogeneous Klein-Gordon equations for ϕ and $\mathbf{A}$ given by

$$\left(\nabla^2 - \frac{1}{c_0^2}\frac{\partial^2}{\partial t^2}\right)\phi(\mathbf{r},t) - \kappa^2\phi = -\frac{\rho}{\epsilon_0}$$

and

$$\left(\nabla^2 - \frac{1}{c_0^2}\frac{\partial^2}{\partial t^2}\right)\mathbf{A}(\mathbf{r},t) - \kappa^2\mathbf{A} = -\mu_0\mathbf{j}$$

respectively. The modifications required to do this yield the Proca equations given by

$$\nabla\cdot\mathbf{E} = \frac{\rho}{\epsilon_0} - \kappa^2\phi, \quad \nabla\cdot\mathbf{B} = 0$$

$$\nabla\times\mathbf{E} = -\frac{\partial\mathbf{B}}{\partial t}, \quad \nabla\times\mathbf{B} = \mu_0\mathbf{j} + \epsilon_0\mu_0\frac{\partial\mathbf{E}}{\partial t} + \kappa^2\mathbf{A}$$

where

$$\mathbf{B} = \nabla\times\mathbf{A}, \text{ and } \mathbf{E} = -\nabla\phi - \frac{\partial\mathbf{A}}{\partial t}.$$

Note that the Klein-Gordon equations for ϕ and $\mathbf{A}$ imply that ϕ and $\mathbf{A}$ and thus $\mathbf{E}$ and $\mathbf{B}$ are effected by mass.

The Proca equations are relativistic field equations that describe massive electromagnetic fields or massive photons (spin 1 vector bosons). They form the foundations for the electro-weak theory (the unification of electromagnetism with the 'weak' force) where it is assumed that the electromagnetic fields of the early universe had significantly greater (relativistic) energies than now, i.e. the electromagnetic and the weak force are manifestation of the same force at relativistic energies. Vector Bosons ($W^{\pm}$ and Z_0 bosons) are taken to be mediators of the weak interaction. However, the Proca equations, as a description for massive photons, have a number of other implications. These include variations in light speed, the possibility of charged black holes, the existence of magnetic monopoles and superluminal (faster than light) particles (Tachyons) with an imaginary mass that can be described by a Proca field with a negative square mass [69], [70] and [71].

The principle associated with deriving the Proca equations can be applied to other field equations such as the Einstein equations for a gravitational field. The Proca-Einstein equations have been used as a basis for modelling the interaction of gravitational fields with dark matter, for example [72]. In string theory, there is tentative evidence that non-Riemannian models such as the Einstein-Proca-Wyle equations may account for dark matter [73]. However, in the context of this thesis, the Proca equations are an example of the modification and extension of a set of field equations in order that a given wave equation is satisfied. Thus, in the derivation of

the Proca equations, the wavefield U is the governing function and not the fields $\mathbf{E}$ and $\mathbf{B}$. In other words, the Proca equations are based on 'tailoring' a field to 'fit' a wavefield. This leads us to consider an approach in which unification is attempted, not in terms of a unified field theory but in terms of a unified wavefield theory where a wavefield U is not just the governing function but the governing principle.

If a unified field theory (unifying gravity and electromagnetism, for example) were available, then, by induction, we should expect that the unifying field equations yield a unifying wave equation. Since a unified field theory is not currently available, our approach is to attempt to construct a unified wavefield theory in which a field is the product of certain characteristics of a wavefield. Thus, the basic idea is to develop a universal physical model that is based on a wavefield equation alone and attempt to explain the characteristics of a field from the wavefield. In this chapter, we adopt the (inhomogeneous) Helmholtz equation and study some of its properties over a broad frequency band including the case when the wavelength approaches infinity. We show how this approach can, for example, be used to explain phenomena such as the 'diffraction' of light by a field that we interpret to be a gravitational field.

9.4 The Inhomogeneous Helmholtz Equation

The three-dimensional inhomogeneous scalar Helmholtz equation can be derived from the (inhomogeneous) time dependent wave equation

$$\left(\nabla^2 - \frac{1}{c^2}\frac{\partial^2}{\partial t^2}\right) U(\mathbf{r}, t) = 0$$

by letting

$$\frac{1}{c^2} = \frac{1}{c_0^2}(1 + \gamma)$$

where $\gamma(\mathbf{r})$ is a dimensionless quantity (the scattering function) and U is a time-dependent scalar wavefield (which is also taken to be dimensionless). We make no demands on the physical nature of U or γ.

With

$$U(\mathbf{r}, t) = u(\mathbf{r}, \omega) \exp(i\omega t)$$

for constant ω (the angular frequency), or with

$$U(\mathbf{r}, t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} u(\mathbf{r}, \omega) \exp(i\omega t) d\omega$$

for variable ω, we obtain the inhomogeneous Helmholtz equation in the form

$$(\nabla^2 + k^2) u(\mathbf{r}, k) = -k^2 \gamma(\mathbf{r}) u(\mathbf{r}, k)$$

where k $(= 2\pi/\lambda)$ is given by

$$k = \frac{\omega}{c_0}.$$

We consider a scattering function γ which is of compact support, i.e.

$$\gamma(\mathbf{r}) \;\; \exists \;\; \forall \;\; \mathbf{r} \in V$$

where V is an arbitrary volume. In electromagnetism, for example, the Helmholtz equation can be derived by decoupling Maxwell's (macroscopic) equations where u describes the scalar electric field and the scattering function is given by $\gamma = \epsilon_r - 1$ where $\epsilon_r \geq 1$ is the isotropic relative permittivity, the relative permeability being taken to be 1 and the conductivity being taken to be zero [74].

9.5 Green's Function Solution for an Incident Plane Wave

Using Green's theorem, the general solution to the inhomogeneous Helmholtz equation at a point $\mathbf{r}_0$ is given by [60], [74],

$$u(\mathbf{r}_0, k) = \oint_S (g\nabla u - u\nabla g) \cdot \hat{\mathbf{n}} d^2\mathbf{r} + k^2 \int_V g\gamma u d^3\mathbf{r}$$

where g is the 'outgoing free space' Green's function given by [74], [75]

$$g(\mathbf{r} \mid \mathbf{r}_0, k) = \frac{\exp(ik \mid \mathbf{r} - \mathbf{r}_0 \mid)}{4\pi \mid \mathbf{r} - \mathbf{r}_0 \mid}$$

which is a solution to the equation

$$(\nabla^2 + k^2) g(\mathbf{r} \mid \mathbf{r}_0, k) = -\delta^3(\mathbf{r} - \mathbf{r}_0)$$

where δ^3 denotes the three-dimensional delta function. Here, S denotes the (closed) surface of the scattering function γ with volume V and $\hat{\mathbf{n}}$ is a unit vector that is perpendicular to an element of the surface $d^2\mathbf{r}$. Note that

$$g(\mathbf{r} \mid \mathbf{r}_0, k) = \frac{1}{4\pi \mid \mathbf{r} - \mathbf{r}_0 \mid}, \;\; k \to 0$$

and thus,

$$\nabla^2 \left(\frac{1}{4\pi \mid \mathbf{r} - \mathbf{r}_0 \mid} \right) = -\delta^3(\mathbf{r} - \mathbf{r}_0).$$

To compute the surface integral, a condition for the behaviour of u on the surface S of γ must be chosen. We consider the case where a simple plane wave of unit amplitude given by

$$u_i(\mathbf{r}, k) = \exp(ik\hat{\mathbf{n}}_i \cdot \mathbf{r})$$

and satisfying the homogeneous Helmholtz equation

$$(\nabla^2 + k^2) u_i(\mathbf{r}, k) = 0$$

is incident on the surface of the scatterer. In this case,

$$u(\mathbf{r}, k) = u_i(\mathbf{r}, k), \;\; \forall \mathbf{r} \in S$$

and we therefore obtain

$$u(\mathbf{r}_0,k) = \oint_S (g\nabla u_i - u_i\nabla g)\cdot\hat{\mathbf{n}}d^2\mathbf{r}k^2 \int_V g\gamma u d^3\mathbf{r} = u_i + u_s$$

where

$$u_s = k^2 \int_V g\gamma u d^3\mathbf{r}.$$

The function u_s is the scattered wavefield which we shall write in the form

$$u_s(\mathbf{r},k) = k^2 g(r,k) \otimes_3 \gamma(\mathbf{r})u(\mathbf{r},k), \quad r =\mid \mathbf{r} \mid$$

where $\otimes_3$ denotes the three-dimensional convolution integral.

9.6 Evaluation of the Scattered Field

To evaluate the scattered field (i.e. to compute u_s), we must define u inside the volume integral. Unlike the surface integral, a boundary condition will not help here because it is not sufficient to specify the behaviour of u at a boundary. In this case, the behaviour of u throughout V needs to be known. This requires a model to be chosen for u inside V that is compatible with a particular physical problem. The simplest model for the internal field is based on assuming that $u \sim u_i \forall \mathbf{r} \in V$. The scattered field is then given by

$$u_s(\mathbf{r}_0,k) = k^2 g(r,k) \otimes_3 \gamma(\mathbf{r})u_i(\mathbf{r},k).$$

This assumption - known as the Born approximation - provides an approximate solution for the scattered field which is valid if

$$k^2 \| g(r,k) \otimes_3 \gamma(\mathbf{r}) \| << 1.$$

This result can be considered to be a first approximation to the (Born) series solution given by

$$u_s(\mathbf{r},k) = u_i(\mathbf{r},k) + k^2 g(r,k) \otimes_3 \gamma(\mathbf{r})u_i(\mathbf{r},k)$$

$$+k^4 g(r,k) \otimes_3 \gamma(\mathbf{r})[g(r) \otimes_3 \gamma(\mathbf{r})u_i(\mathbf{r},k)] + ...$$

which is valid under the condition

$$k^2 \| g(r,k) \otimes_3 \gamma(\mathbf{r}) \| < 1.$$

Each term in this series expresses the effects due to single, double and triple etc. scattering events. Because this series scales as k^2, k^4, k^6, ..., for a fixed $k << 1$ (long wavelength wavefields), the Born approximation becomes an exact solution.

9.7 Low Frequency Helmholtz Scattering

If a Helmholtz wavefield oscillates at lower and lower frequencies, then we can consider an asymptotic solution of the form

$$u_s(\mathbf{r}_0, k) = \frac{k^2}{4\pi} \int_V \frac{\gamma(\mathbf{r})}{\mid \mathbf{r} - \mathbf{r}_0 \mid} u_i(\mathbf{r}, k) d^3\mathbf{r}, \quad k \to 0.$$

This is a consequence of the fact that the higher order terms in the Born series can be ignored leaving just the first term as $k \to 0$ and because

$$\frac{\exp(ik \mid \mathbf{r} - \mathbf{r}_0 \mid)}{4\pi \mid \mathbf{r} - \mathbf{r}_0 \mid} = \frac{1}{4\pi \mid \mathbf{r} - \mathbf{r}_0 \mid}, \quad k \to 0$$

giving an exact solution to the problem.

If the incident field is a unit plane wave, then

$$u(\mathbf{r}_0, k) = 1 + u_s(\mathbf{r}_0, k)$$

where

$$u_s(\mathbf{r}_0, k) = \frac{k^2}{4\pi} \int_V \frac{\gamma(\mathbf{r})}{\mid \mathbf{r} - \mathbf{r}_0 \mid} d^3\mathbf{r}, \quad k \to 0$$

which we write in the form

$$u_s(\mathbf{r}, k) = \frac{k^2}{4\pi r} \otimes_3 \gamma(\mathbf{r}), \quad k \to 0.$$

Here, the wavelength of the incident plane wavefield is assumed to be significantly larger than the spatial extent V of the scatterer. For a given scattering function $\gamma(\mathbf{r})$ the wavefield is a 'weak field' because of the low values of k required to produce this (asymptotic) result. But this result is the general solution to Poisson's equation

$$\nabla^2 u_s(\mathbf{r}, k) = -k^2 \gamma(\mathbf{r})$$

since, using the result

$$\nabla^2 \left(\frac{1}{4\pi r} \right) = -\delta^3$$

we have

$$\nabla^2 u = \nabla^2 u_s = k^2 \nabla^2 \left(\frac{1}{4\pi r} \otimes_3 \gamma \right)$$

$$= k^2 \gamma \otimes_3 \nabla^2 \left(\frac{1}{4\pi r} \right) = -k^2 \gamma \otimes_3 \delta^3 = -k^2 \gamma.$$

By considering u_s to be a potential, we can write

$$\nabla \cdot \mathbf{U}_s(\mathbf{r}, k) = k^2 \gamma(\mathbf{r}), \quad \mathbf{U}_s(\mathbf{r}, k) = -\nabla u_s(\mathbf{r}, k).$$

Integrating over the volume of the scatterer V, we obtain

$$\int_V \nabla \cdot \mathbf{U}_s(\mathbf{r},k)d^3\mathbf{r} = k^2 \int_V \gamma(\mathbf{r})d^3\mathbf{r}$$

and using the divergence theorem we can write

$$\oint_S \mathbf{U}_s(\mathbf{r},k)\cdot\hat{\mathbf{n}}d^2\mathbf{r} = k^2\Gamma, \quad \Gamma = \int_V \gamma(\mathbf{r})d^3\mathbf{r}.$$

If we now consider a scatterer that is a sphere, then the field $\mathbf{U}$ will have radial symmetry, i.e. $\mathbf{U}_s = \hat{\mathbf{n}}U_s$. In this case, the surface integral becomes $4\pi r^2 U_s$ and we obtain

$$U_s = \frac{k^2\Gamma}{4\pi r^2}, \quad k \to 0.$$

Hence, in the limit as $k \to 0$, Helmholtz scattering provides an exact solution for a weak field whose gradient (for the radially symmetric case) is characterized by a $1/r^2$ scaling law.

9.8 Diffraction

For $k \to 0$, $u_s(\mathbf{r},k)$, which we now denote by $u_s^0(\mathbf{r},k_0)$, is the solution to

$$\nabla^2 u_s^0(\mathbf{r},k_0) = -k_0^2\gamma(\mathbf{r})$$

where k_0 denotes a value for k, $k \to 0$. Consider a Born scattered Helmholtz wavefield $u_s(\mathbf{r},k)$ for $k >> 1$ given by

$$u_s(\mathbf{r},k) = k^2 g(r,k) \otimes_3 \gamma(\mathbf{r})u_i(\mathbf{r},k).$$

We can then write

$$u_s(\mathbf{r},k) = -\frac{k^2}{k_0^2} g(r,k) \otimes_3 u_i(\mathbf{r},k)[\nabla^2 u_s^0(\mathbf{r},k_0)]$$

from which we can derive an expression for the far field scattering amplitude generated by the field $\mathbf{U}_s^0$ given by

$$u_s(\mathbf{r},k) = -\frac{k^2}{k_0^2} g(r,k) \otimes_3 u_i(\mathbf{r},k)[\nabla \cdot \mathbf{U}_s^0(\mathbf{r},k_0)]$$

$$= \frac{\exp(ikr_0)}{4\pi r_0} A(\hat{\mathbf{n}}_0, \hat{\mathbf{n}}_i), \quad \frac{r}{r_0} << 1$$

where, with $u_i(\mathbf{r},k) = \exp(ik\hat{\mathbf{n}}_i \cdot \mathbf{r})$, $\hat{\mathbf{n}}_0 = \mathbf{r}_0 / \mid \mathbf{r}_0 \mid$ and

$$\mathbf{U}_s^0 = \hat{\mathbf{n}}U_s^0 = \hat{\mathbf{n}}\frac{k_0^2\Gamma}{4\pi r^2},$$

$$A(\hat{\mathbf{n}}_0, \hat{\mathbf{n}}_i) = -\frac{k^2\Gamma}{4\pi}\int_V \exp[-ik(\hat{\mathbf{n}}_0 - \hat{\mathbf{n}}_i)\cdot\mathbf{r}]\nabla\cdot\left(\frac{\hat{\mathbf{n}}}{r^2}\right)d^3\mathbf{r}.$$

Hence, the wavefield $u_s(\mathbf{r}, k)$ (for $k >> 1$) generated by a scatterer that is simultaneously generating a scattered wavefield $u_s^0(\mathbf{r}, k_0)$ is, in the far field (under the Born approximation) determined by the Fourier transform of the scattering function (assuming radial symmetry) $f(r) = \nabla\cdot(\hat{\mathbf{n}}r^{-2})$. In other words, the weak field generated by very low frequency scattering will diffract a high frequency Helmholtz wavefield, the diffraction pattern (i.e. the far field scattering pattern) being determined by $f(r)$. This is an example of a low frequency Helmholtz scattered field scattering a high frequency Helmholtz field, the field being of the same type but characterised by a (very) large difference in frequency.

9.8.1 Diffraction by an Infinitely Thin Scatterer

Consider the case where an incident plane wavefield is travelling in the z-direction, i.e. $u_i = \exp(ikz)$ and is incident on an infinitely thin scatterer defined by the function $\gamma(\mathbf{r}) = \gamma(x, y)\delta(z)$. The scattered wavefield is then given by

$$u_s(x, y, z, k)$$

$$= k^2\frac{\exp(ik\sqrt{x^2+y^2+z^2})}{4\pi\sqrt{x^2+y^2+z^2}} \otimes_3 \gamma(x, y)\delta(z)\exp(ikz)$$

$$= k^2\frac{\exp(ik\sqrt{x^2+y^2+z^2}}{4\pi\sqrt{x^2+y^2+z^2}} \otimes_2 \gamma(x, y), \quad \gamma \ \exists \ \forall(x, y) \in S$$

where $\otimes_2$ denotes the two-dimensional convolution integral over area S. Writing out this result in the form

$$u_s(x_0, y_0, z_0, k)$$

$$= k^2\int\int \frac{\exp[ik\sqrt{(x-x_0)^2+(y-y_0)^2+z_0^2}]}{4\pi\sqrt{(x-x_0)^2+(y-y_0)^2+z_0^2}}\gamma(x, y)dxdy,$$

it is clear that if the scattered wavefield is now measured in the far field, i.e. for the case when $x/z_0 << 1$ and $y/z_0 << 1$, then

$$z_0\left(1 + \frac{(x-x_0)^2}{z_0^2} + \frac{(y-y_0)^2}{z_0^2}\right)^{\frac{1}{2}}$$

$$\simeq z_0 - \frac{xx_0}{z_0} - \frac{yy_0}{z_0} + \frac{x_0^2}{2z_0} + \frac{y_0^2}{2z_0}$$

and thus,

$$u_s(x_0, y_0, z_0, k) = \frac{\exp(ikz_0)}{4\pi z_0}\exp\left(ik\frac{x_0^2+y_0^2}{2z_0}\right)A(k_x, k_y)$$

where

$$A(k_x, k_y) = k^2 \widetilde{\gamma}(k_x, k_y) = k^2 \mathcal{F}_2[\gamma(x, y)]$$

$$= k^2 \int \int \exp(-iux) \exp(-ivy) \gamma(x, y) dx dy$$

with spatial frequencies u and v being defined by

$$u = \frac{kx_0}{z_0} = \frac{2\pi x_0}{\lambda z_0}$$

and

$$v = \frac{ky_0}{z_0} = \frac{2\pi y_0}{\lambda z_0}.$$

Here, $\mathcal{F}_2$ denotes the two-dimensional Fourier transform, the result being the standard expression for a diffraction pattern in the far field or Fraunhofer zone [74].

9.8.2 Diffraction by an Infinitely Thin Field

In the previous section, we derived the far field diffraction pattern for an infinitely thin scatterer. However, suppose this scatterer also radiates a field generated by low frequency Helmholtz scattering from the same scattering function. What is the contribution of this field to the diffraction of the same incident plane wave within and beyond the extent of the scatterer[1]? In this case, the scattered wavefield is given by (under the Born approximation)

$$u_s = -\frac{k^2}{k_0^2} g \otimes_3 u_i \nabla^2 u_s^0, \quad u_s^0 = \frac{k_0^2}{4\pi r} \otimes_3 \gamma.$$

For an infinitely thin scatterer given by $\gamma(x, y)\delta(z)$,

$$u_s^0(x, y, z, k_0) = \frac{k_0^2}{4\pi \sqrt{x^2 + y^2 + z^2}} \otimes_2 \gamma(x, y)$$

so that in the (x, y) plane located at $z = 0$,

$$u_s^0(x, y, k_0) = \frac{k_0^2}{4\pi \sqrt{x^2 + y^2}} \otimes_2 \gamma(x, y).$$

For an incident plane wave $u_i = \exp(ikz)$, the scattered wavefield u_s is thus, given by

$$u_s(x, y, z, k) = -k^2 g(r, k) \otimes_3 \exp(ikz) ...$$

$$\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} \right) \left(\frac{1}{4\pi \sqrt{x^2 + y^2}} \otimes_2 \gamma(x, y) \right).$$

[1] Note that the scattered wavefield u_s^0 is taken to exist within and beyond the finite spatial extent of the scatterer $\gamma(\mathbf{r}), \mathbf{r} \in V$, i.e. u_s^0 is not of compact support since it is given by the convolution of a function of compact support with r^{-1}.

Repeating the calculation given in the previous section (for $z \to 0$), the diffracted wavefield now becomes

$$u_s(x_0, y_0, z_0, k) = \frac{\exp(ikz_0)}{4\pi z_0} \exp\left(ik\frac{x_0^2 + y_0^2}{2z_0}\right) A(k_x, k_y)$$

where

$$A(k_x, k_y) = -2zk^2 \mathcal{F}_2 \left[\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right) \frac{1}{4\pi\sqrt{x^2 + y^2}} \otimes_2 \gamma(x, y)\right].$$

Note that although the scatterer is taken to be 'infinitely thin' because $\gamma(\mathbf{r}) = \gamma(x, y)\delta(z)$, we still consider the physical thickness of the scatterer to be finite[2], i.e. $z \neq 0$. Now, for an arbitrary function $f \Longleftrightarrow \widetilde{f}$, where $\Longleftrightarrow$ denotes the transform from real space to Fourier space [74],

$$\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right) f \Longleftrightarrow -(k_x^2 + k_y^2)\widetilde{f},$$

$$\frac{1}{\sqrt{x^2 + y^2}} \Longleftrightarrow \frac{2\pi}{\sqrt{u^2 + v^2}},$$

and we obtain

$$A(k_x, k_y) = zk^2\sqrt{u^2 + v^2}\widetilde{\gamma}(k_x, k_y).$$

Figure 35 shows numerical simulations of the diffraction patterns compounded in the (intensity) functions

$$\mid \widetilde{\gamma}(k_x, k_y) \mid^2 \quad \text{and} \quad u^2 + v^2 \mid \widetilde{\gamma}(k_x, k_y) \mid^2$$

using a two-dimensional Fast Fourier Transform for the case when the scattering function is given by the rotationally symmetric functions (for $r = \sqrt{x^2 + y^2}$)

$$\gamma(r) = \exp(-r^2/\sigma^2)$$

(a unit amplitude Gaussian function[3] with standard deviation σ) and (a unit amplitude disc function)

$$\gamma(r) = \begin{cases} 1, & r \leq a; \\ 0, & \text{otherwise.} \end{cases}$$

The analytical solutions, for the intensity

$$I_1 = \mid u_s \mid^2$$

generated by diffraction from the scatterer γ and

$$I_2 = \mid u_s \mid^2$$

---

2 z should be taken to be a positive real 'infinitesimal' for all real k.

3 Taken by default, to be of finite extent.

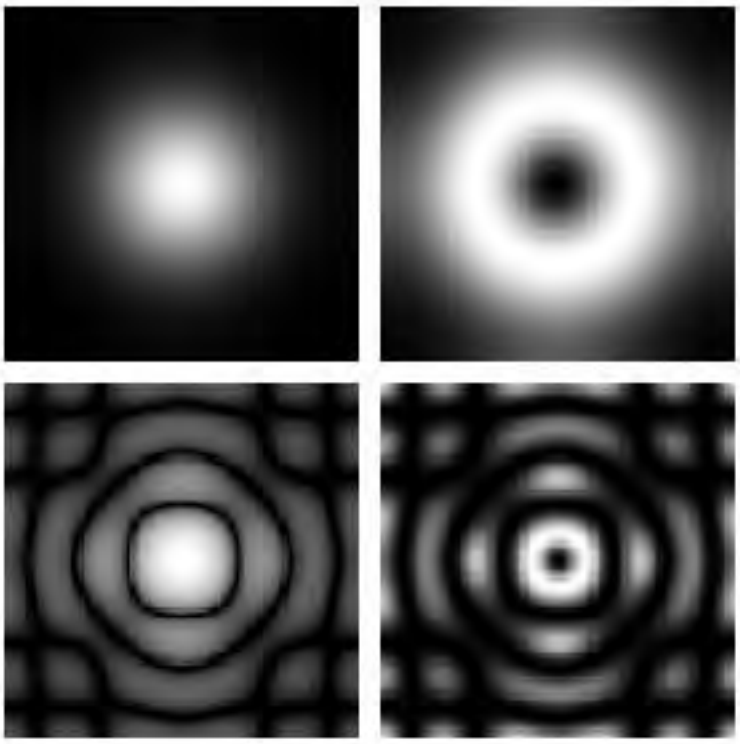

FIGURE 35 Numerical simulation of the intensity patterns for an Gaussian function (top) and disc function (bottom) associated with the diffraction of a wave-field by an infinitely thin scatterer $\gamma(x, y)$ (left - plotted using a logarithmic scale) and the field $\nabla^2 u_s^0$ generated by the same scatterer.

generated by the field $\nabla^2 u_s^0$ are given by:

$$I_1(r_0,\lambda) = \frac{\pi^4\sigma^2}{z_0^2\lambda^4}\exp\left[-\left(\frac{2\pi^2\sigma^2 r_0^2}{\lambda^2 z_0^2}\right)\right]$$

and

$$I_2(r_0,\lambda) = z^2\frac{4\pi^6\sigma^2 r_0^2}{z_0^4\lambda^6}\exp\left[-\left(\frac{2\pi^2\sigma^2 r_0^2}{\lambda^2 z_0^2}\right)\right]$$

for a Gaussian diffractor and, for a disc diffractor, with $\xi = \frac{2\pi a r_0}{\lambda z_0}$,

$$I_1(r_0,\lambda) = \frac{4\pi^4 a^4}{z_0^2\lambda^4}\left(\frac{J_1(\xi)}{\xi}\right)^2$$

and

$$I_2(r_0,\lambda) = z^2\frac{16\pi^6 a^4 r_0^2}{z_0^4\lambda^6}\left(\frac{J_1(\xi)}{\xi}\right)^2.$$

Note that the Gaussian ring has a maximum when $r_0 = z_0\lambda/(\sqrt{2}\pi\sigma)$ and that, in the latter case, the diffraction pattern is determined by the 'jinc' function $J_1(\xi)/\xi$ whose first minimum occurs when $\xi = 3.83$, i.e. when

$$r_{\min} = 1.22\frac{\lambda z_0}{a}$$

which is a classical result in (Fourier) optics - an Airy pattern [74]. Observe that the magnitude of the intensity patterns generated by the field $\nabla^2 u_s^0$ is significantly less than the scatterer γ, e.g. in the case of a Gaussian function

$$\frac{I_2}{I_1} = \frac{4z^2\pi^2 r_0^2}{z_0^2\lambda^2}$$

and only if $r_0/\lambda \sim z/z_0$ will the magnitude become of the same order. Also observe that the intensity generated by the scatterer γ scales as λ^{-4} whereas the intensity generated by the field $\nabla^2 u_s^0$ scales as λ^{-6}. However, the most significant result is that diffraction for a scattering function produces a pattern whose intensity peaks at the centre of the image plane (a standard result in Fourier optics) but that diffraction from a low frequency scattered field produces a pattern characterised by a ring. The multiplicity of rings in either case is determined by whether or not the scattering function is discontinuous.

9.9 The Poisson Spot and the Einstein Ring

Consider the images given in Figure 36 which show an example of a Poisson (or Arago) spot [76] and an Einstein ring [77]. The Poisson spot (named after Simeon Poisson who investigated the phenomenon in 1818) represents a landmark in the

history of science in terms of validating whether or not light was a particle or a wave. The Poisson spot is a bright compact feature (a spot) that appears at the centre of the shadow of a circular opaque object. In Figure 36, the Poisson spot is the result of laser light diffracting from the edge of a ball-bearing. In a theoretical model of this effect, the ball-bearing can be replaced by an infinitely thin disc. However, because this disc is opaque, the scattering function must be defined by

$$\gamma(r) = \begin{cases} 0, & r \leq a; \\ 1, & \text{otherwise.} \end{cases}$$

and the Fourier transform (assuming an incident plane wave $\exp(ikz)$ that is of infinite extent over the (x, y) plane) must be taken from $-\infty$ to $-a$ and from a to ∞. This is equivalent to computing the two-dimensional Fourier transform over all space and subtracting the Fourier transform over $r \leq a$. Since

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp(-iux)\exp(-ivy)dxdy = 4\pi^2\delta(k_x)\delta(k_y)$$

the diffracted intensity for an opaque object is

$$I_1(r_0, \lambda) = \delta^2(r_0) + \frac{\pi^4 a^4}{z_0^2\lambda^4}\left(\frac{2J_1(\xi)}{\xi}\right)^2.$$

The fact that the Poisson spot occurs within the geometrical shadow of an opaque object, is evidence that a particle and/or a geometrical theory of optics is invalid and that light must therefore be a wavefield. This deduction occurred some forty years before Faraday and Maxwell concluded that light was indeed a wave but one composed of electric and magnetic fields - a direct consequence of the fact that the field equations derived by Maxwell for an electric and magnetic field can be decoupled to yield a wave equation.

9.9.1 Gravitational Diffraction

The Einstein ring shown in Figure 36 is an effect that is conventionally explained in terms of the bending of light through the curvature of space (and time) by a mass. This is a consequence of the field equations for a gravitational field (the Einstein equations [56]). In order to obtain an Einstein ring, the magnitude of the gravitational field must be relatively high such as that generated by a spiral galaxy. Further, in order to generate a near perfect (complete) ring, the entire galaxy must be well aligned with regard to an observer in the 'object plane'. The bending of light by a gravitational field has an analogy with the geometrical interpretation of light interacting with a lens. At the edge of a lens, the light beam is 'bent' (discontinuously) by the change in refractive index from air to glass and from glass to air - the extreme edge of a lens acts like a prism. Like an optical lens, gravitational 'lensing' will produce distortions of the object plane when alignment of the 'earth-lens-object' is imperfect.

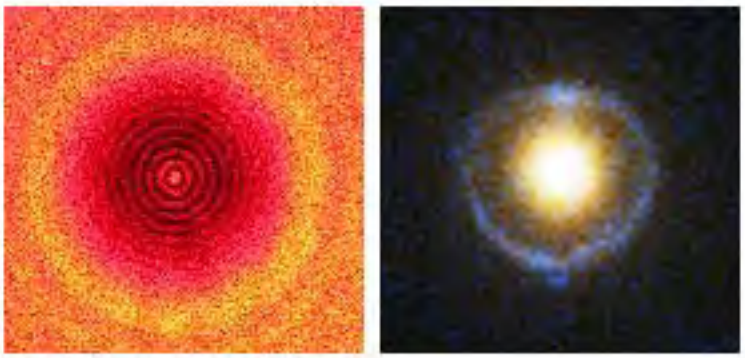

FIGURE 36 Diffraction pattern from the incidence of laser light with a ball-bearing illustrating the Poisson spot (left) and an example of an Einstein ring generated by a spiral galaxy (central feature) observed with the Hubble Space Telescope (right).

If we interpret an Einstein ring in terms of the results given in Section 9.8.2, then the ring is not due to light being bent (continuously) by the curvature of a space-time continuum but the result of the diffraction of a plane wave (i.e. light) by the field $\nabla^2 u_s^0$ which is taken to be in the plane of the galaxy and to extend beyond it. This requires the magnitude of the scattering function to be very large in order to compensate for $z \to 0$. If we model a (spiral) galaxy in terms of a Gaussian function, then the ring associated with the diffraction pattern given in Figure 35 is, in this sense, a simulation of the Einstein ring given in Figure 36. The use of a Gaussian function to model the macroscopic gravitational field generated by a spiral galaxy is intuitive as the edges of a galaxy will not be discontinuous (especially on the scale of the wavelength of light!). However, in the case of a black hole, the event horizon defines an edge. In such a case, we can expect gravitational diffraction to produce a number of concentric rings similar to those associated with a Poisson spot, the black hole being modelled in terms of an opaque disc. Multiple ring patterns associated with a black hole are a prediction of the conventional bending of light by space-time curvature. The idea is that, close to the event horizon, the gravitational field is so intense that light can be curved right around the black hole by 180 degrees or more to produce a ring associated with the light generated by an object that exists in alignment with, and behind, the image plane [77]. These multiple Einstein ring predictions are based on arguments analogous to geometric optics whereas the multiple rings considered here are analogous to Fourier optics. In this sense, we are interpreting a gravitational field to be generated by the scattering of a long wavelength Helmholtz wavefield, i.e. the field $\mathbf{U}_s^0$ defines a 'gravitational field'.

9.9.2 Colour Analysis

Another feature of Einstein rings (complete or otherwise) is that, unless the source-galaxy system has been substantially red shifted (when both the galaxy and the ring appear red, e.g. [78]), the colour of the rings is blue (as in the example given in

Figure 36) even, as in some cases that have been reported, when the galaxy itself is red [79]. Note that there are many examples of Einstein rings in the infrared and radio spectra which are not blue due to the false colour mappings that are used to display the data. However, it the visible spectrum, the colour of an Einstein ring is blue which is the true colour of the effect as observed by the Hubble Space Telescope, for example. The issue of why an Einstein ring is blue does not appear to have received musch attention and a variety of of unacceptable explanation have been provided. For example, in the Astronomy Picture the Day of July 28 2008 [80] a NASA image of a Galaxy Einstein Ring is accompanied by the following: *What's large and blue and can wrap itself around an entire galaxy? A gravitational lens image. Pictured above on the left, the gravity of a normal white galaxy has gravitationally distorted the light from a much more distant blue galaxy.* It is clearly not conceivable that all Einstein rings observed to date (in the visible spectrum) are the result of gravitational lensing of light from blue galaxies. Galaxies are not blue anyway but a source of radiation over the entire electromagnetic spectrum.

If we accept an Einstein ring to be a gravitational diffraction phenomena, then the intensity of the diffracted light scales as λ^{-6} which explains the colour of the rings (blue light having the shortest wavelength in the visible spectrum). This is analogous to the explanation of why the Earth's atmosphere is blue in colour. Under the Rayleigh scattering condition in which the wavelength is significantly larger than the physical size of the scatterer (when the Born approximation is valid), the scattering amplitude becomes independent of the scattering angle and the intensity of the scattered field is proportional to λ^{-4}. Thus, the sky is blue, because sunlight is scattered by the electrons of air molecules in the terrestrial atmosphere generating blue light preferentially around in all directions. Further, as the Sun approaches the horizon, we have to look more and more diagonally through the Earth's atmosphere. Our line of sight through the atmosphere is then longer and most of the blue light is scattered out before it reaches us, especially as the Sun gets very near the horizon. Relatively more red light reaches us, accounting for the reddish colour of sunsets. In other words, the λ^{-4} dependence of the scattered intensity implies that the atmosphere scatters green, blue and violet light photons more effectively than yellow, orange, and red photons. As the Sun approaches the horizon, the path of light through the atmosphere increases, so more of the short-wavelength photons get scattered away leaving the longer-wavelength photons and the Sun looks progressively redder. Rayleigh scattering in the atmosphere also explains why the sun is yellow at mid-day. This is because the energy spectrum (i.e. Planck's radiation law [74]) for the Sun peaks at the point when the wavelength is that of green light (i.e. $\sim 4.7 \times 10^{-7}$metres). Since the atmosphere filters out blue light and since blue and yellow light combine to give green light, the Sun appears yellow.

Note that the λ^{-6} scaling dependency associated with gravitational diffraction provides a method of validating or otherwise the theoretical model presented here. We require a scenario in which the same Einstein ring is recorded simultaneously over a broad frequency spectrum (e.g. using radio, infrared, visible and ultraviolet imaging) in such a way that the intensities of each image (relative to a known source that can be used for calibration) can be compared on a quantitative basis.

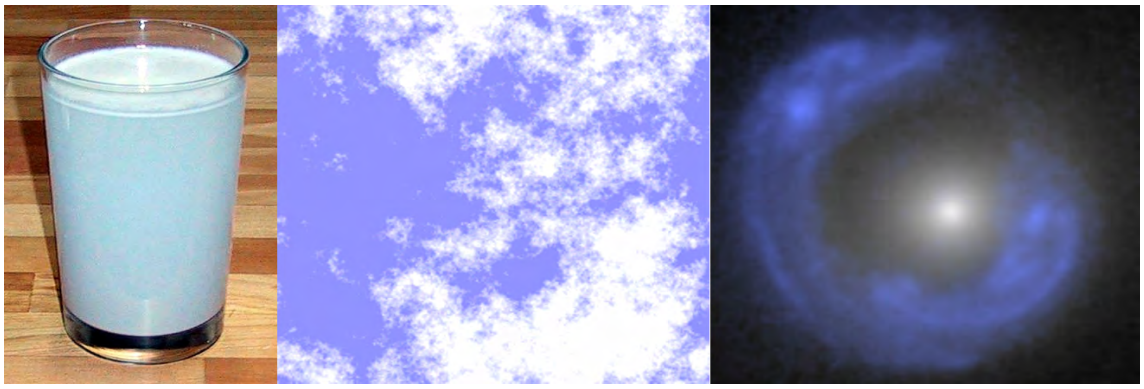

FIGURE 37 Examples of the differences in the lightness (for a HSL - Hue, Saturation and Lightness - colour model) of the blue light generated by Tyndall scattering (scattering of light by fine flour suspended in water - left), Rayleigh scattering (scattering of light by the atmosphere - centre) and 'gravitational scattering' (diffraction of light by the gravitation field generated by a galaxy - right).

However, data available to undertake such an analysis are not yet available. Instead, another approach is considered based on the colour generated by light scattered under different conditions. For Tyndall scattering discussed in Section 4.3.3, the intensity of light is proportional to λ^{-2} and for Rayleigh scattering discussed in Section 4.3.2, the light scattered intensity is proportional to λ^{-4}. Because of these wavelength scaling relationships, both Tyndall and Rayleigh scattering generate blue light. However, Tyndall scattering can be expected to generate a lighter blue than Rayleigh scattering. This is illustrated in Figure 37 which also shows, for comparison, the colour of the blue light scattered by a gravitational field which is proportionately darker because of the scaling relationship characterised by λ^{-6}. This comparison is quantified in Figure 38 which shows the differences in the lightness of blue using a Hue, Saturation and Lightness (HSL) colour model. The lightness factor associated with each image is characterised by the ratios 1:2:3 which is in agreement with the logarithmic scaling ratios $2\ln\lambda : 4\ln\lambda : 6\ln\lambda$.

9.10 Schrödinger Scattering

The theoretical ideas established so far and some of the implications that have been discussed are without reference to any physical significance of the scattering function. In this section (and the following section) we examine the characteristics of this scattering function by revisiting two wave equations in quantum mechanics, namely the Schrödinger equation (for the non-relativistic case) and the Klein-Gordon equation (for the relativistic case).

If we consider the diffraction of light by a material object, then physically, the scattering function $\gamma(\mathbf{r})$ must describe some appropriate property of matter (the

FIGURE 38 The 'blues' associated with Tyndall (left), Rayleigh (centre) and gravitational (right) light scattering obtained by averaging over many images of each effect.

material properties) that is consistent with electromagnetic theory. On the macroscopic scale (i.e. many orders of wavelength) the relative permittivity, permeability and conductivity are the basis for defining Maxwell's macroscopic equations [74]. These material properties vary considerably from one application to the next. They may be isotropic or non-isotropic functions of space, time varying and field varying (non-linear optics), for example.

In electromagnetism, the use of the scalar Helmholtz equation to develop the results given so far, is compatible only with the case when the relative permeability is 1, the conductivity is zero and when the material is isotropic (i.e. the relative permittivity is a scalar function of space). However, in terms of a universal wavefield theory, matter is ultimately composed of matter waves which conform to matter wave equations such as the Schrödinger equation.

The fundamental postulates of quantum mechanics are that $E = \hbar\omega$ and $\mathbf{p} = \hbar\mathbf{k}$. Given that

$$E = \frac{p^2}{2m}$$

then

$$\frac{1}{c^2} = \frac{k^2}{\omega^2} = \frac{p^2}{E^2} = \frac{2m}{E}$$

and the wave equation

$$\left(\nabla^2 - \frac{1}{c^2}\frac{\partial^2}{\partial t^2}\right) U(\mathbf{r}, t) = 0, \quad \frac{1}{c^2} = \frac{1}{c_0^2}(1+\gamma)$$

can be written in terms of the Helmholtz equation

$$(\nabla^2 + k^2)u(\mathbf{r}, k) = -k^2\gamma u(\mathbf{r}, k), \quad \gamma = \frac{2mc_0^2}{E} - 1.$$

Note that for a potential energy function E_p when

$$E = \frac{p^2}{2m} + E_p,$$

the scattering function is given by

$$\gamma = \frac{2mc_0^2(E - E_p)}{E^2} - 1.$$

In either case, we note that Schrödinger's equation is obtained when the angular frequencies defining k and E are the same. Thus, the scattering function associated with the Helmholtz equation given above is, in this sense, a generalization of Schrödinger's equation where the wavefield $U(\mathbf{r}, t)$ can oscillate at any frequency ω less than, or significantly less than the frequency, ω_1 say, associated with a matter wave of energy $E = \hbar\omega_1$. Schrödinger's equation is therefore taken to be a 'product' of the limiting case: $\omega \to \omega_1$[4].

Defining the scattering function in this way, we note that

$$U_s^0 = \frac{k_0^2\Gamma}{4\pi r^2}$$

where, for constant E and m,

$$\Gamma = Mm$$

and

$$M = \frac{V}{m}\left(\frac{2mc_0^2}{E} - 1\right), \quad V = \int_V d^3\mathbf{r}.$$

Suppose that a mass m', placed in the vicinity of the field U_s^0, experiences a force F that is proportional to Um' so that

$$F = v^2 Um'$$

where v^2 is a constant of proportionality. Then

$$F = v^2k_0^2\frac{\Gamma m'}{4\pi r^2} = G\frac{mm'}{r^2}, \quad G = \frac{Mv^2k_0^2}{4\pi}$$

and v has the dimensions of velocity (i.e. $\mathsf{length.second}^{-1}$). We can then derive an expression for the wavelength of the field U_s^0 in terms of the gravitational constant G, i.e.

$$\lambda_0 = \frac{2\pi}{k_0} = \frac{c_0}{\nu}$$

where ν is the frequency given by

$$\nu = r\frac{c_0}{v^2}\sqrt{\frac{Gm}{\pi V}}, \quad r = \sqrt{\frac{E}{2mc_0^2 - E}} \quad .$$

Note that for the frequency (and wavelength) to be a real positive quantity, we require that

$$2mc_0^2 > E$$

[4] An entirely phenomenological argument (like Schrödinger's equation itself!).

so that

$$\frac{2mc_0^2}{E} - 1 > 0 \implies \gamma > 0.$$

Also note that because v has dimensions of velocity, the 'force field has an associated 'speed'.

The inhomogeneous Helmholtz equation

$$\left(\nabla^2 + \frac{\omega^2}{c_0^2}\right) u = -\frac{\omega^2}{c_0^2}\gamma u$$

where

$$\gamma = 2mc_0^2(E - E_p)/E^2 - 1$$

is the Schrödinger equation in 'disguise' in the sense that if $\omega \to \omega_1$ where $E = \hbar\omega_1$, then

$$(\nabla^2 + k_1^2)u = \gamma_1 u$$

where

$$k_1^2 = \frac{\omega_1^2}{c_0^2} = \frac{2mE}{\hbar^2} \text{ and } \gamma_1 = \frac{2mE_p}{\hbar^2}.$$

Given that Proca's equations can be decoupled to produce inhomogeneous Klein-Gordon equations for ϕ and $\mathbf{A}$, we can adopt the same procedure to obtain the following inhomogeneous wave equations for the non-relativistic case, i.e.

$$\left(\nabla^2 - \frac{1}{c_0^2}\frac{\partial^2}{\partial t^2}\right)\phi(\mathbf{r},t) - \gamma\frac{1}{c_0^2}\frac{\partial^2\phi}{\partial t^2} = -\frac{\rho}{\epsilon_0}$$

and

$$\left(\nabla^2 - \frac{1}{c_0^2}\frac{\partial^2}{\partial t^2}\right)\mathbf{A}(\mathbf{r},t) - \gamma\frac{1}{c_0^2}\frac{\partial^2\mathbf{A}}{\partial t^2} = -\mu_0\mathbf{j},$$

Maxwell's equations being modified to the form

$$\nabla\cdot\mathbf{E} = \frac{\rho}{\epsilon_0} - \gamma\frac{1}{c_0^2}\frac{\partial^2\phi}{\partial t^2}, \quad \nabla\cdot\mathbf{B} = 0$$

$$\nabla\times\mathbf{E} = -\frac{\partial\mathbf{B}}{\partial t}, \quad \nabla\times\mathbf{B} = \mu_0\mathbf{j} + \epsilon_0\mu_0\frac{\partial\mathbf{E}}{\partial t} + \gamma\frac{1}{c_0^2}\frac{\partial^2\mathbf{A}}{\partial t^2}.$$

The fields ϕ_s^0 and $\mathbf{A}_s^0$ (the equivalent of U_s^0) are given by

$$\phi_s^0 = \frac{k_0^2\Gamma}{4\pi r^2} + \frac{P}{4\pi\epsilon_0 r^2}$$

and

$$\mathbf{A}_s^0 = \hat{\mathbf{n}}_0\frac{k_0^2\Gamma}{4\pi r^2} + \frac{\mu_0\mathbf{J}}{4\pi r^2}, \quad \hat{\mathbf{n}}_0 = \mathbf{A}_s^0/\mid\mathbf{A}_s^0\mid$$

where, for time-independent functions ρ and $\mathbf{J}$,

$$P = \int_V \rho(\mathbf{r})d^3\mathbf{r} \text{ and } \mathbf{J} = \int_V \mathbf{j}(\mathbf{r})d^3\mathbf{r}.$$

Note that for the limiting case when $\omega \rightarrow \omega_1$ we obtain modified Schrödinger equations for ϕ and $\mathbf{A}$ given by

$$(\nabla^2 + k_1^2)\phi = \gamma_1 \phi - \frac{\rho}{\epsilon_0}$$

and

$$(\nabla^2 + k_1^2)\mathbf{A} = \gamma_1 \mathbf{A} - \mu_0 \mathbf{j}.$$

In the context of the results above, we interpret the field U_s^0 in terms of a low frequency electric scalar potential (in a charge free environment with $\rho = 0$). In this sense, we could interpret the field U_s^0 as an ultra low frequency electromagnetic field in terms of an answer to the question: how long does a radio wave have to be before it becomes something else? However, in the universal wave model considered here, fields such as ϕ and $\mathbf{A}$ are subservient to the wavefield characterised by a governing wave equation in a similar sense to the rationale associated with the derivation of the Proca equations. Thus, the issue as to whether U_s is interpreted in terms of an electromagnetic, gravitational or quantum field is redundant, at least in the conventional sense. Rather, we consider all fields such as ϕ to be a characteristic of wavefields interacting over a broad frequency range. In this sense, the use of a scalar wavefield U in quantum mechanical equations such as the Schrödinger and Klein-Gordon equations is also being used in the interpretation of electromagnetism and gravitation. Field equations such as Maxwell's and Einstein equation's must be re-interpreted and derived from a universal wavefield approach alone, along with the physical interpretation of an electric and gravitational field.

9.11 Klein-Gordon Scattering

For the relativistic case

$$E^2 = p^2 c_0^2 + m^2 c_0^4$$

and

$$\frac{1}{c^2} = \frac{k^2}{\omega^2} = \frac{p^2}{E^2} = \frac{1}{c_0^2} - \frac{m^2 c_0^2}{E^2}.$$

The wave equation

$$\left(\nabla^2 - \frac{1}{c^2}\frac{\partial^2}{\partial t^2}\right) U(\mathbf{r}, t) = 0$$

can thus be written in terms of the Helmholtz equation as

$$(\nabla^2 + k^2)u(\mathbf{r}, k) = -k^2 \gamma u(\mathbf{r}, k)$$

where γ is the 'Klein-Gordon scattering function' given by

$$\gamma = -\frac{m^2 c_0^4}{E^2}$$

The field U_s^0 is then given by

$$U_s^0 = -\frac{k_0^2 \Gamma}{4\pi r^2}$$

where (for constant E and m)

$$\Gamma = Mm^2, \;\; M = \frac{c_0^4 V}{E^2}.$$

We note that in this case, U_s^0 is proportional to the square of the mass and is of negative polarity compared to the non-relativistic case, i.e. it will generate a repulsive force on a particle of mass m' given by

$$F = -G\frac{m^2 m'}{r^2}.$$

9.12 Intermediate Scattering

Since (for positive energies)

$$E = \sqrt{p^2 c_0^2 + m^2 c_0^4} \simeq \frac{p^2}{2m} + mc_0^2, \;\; \frac{p^2}{m^2 c_0^2} << 1$$

we recover Schrödinger's equation

$$i\hbar \frac{\partial U}{\partial t} = -\frac{\hbar^2}{2m}\nabla^2 U + mc_0^2 U$$

which now includes the rest mass energy term $mc_0^2 U$. In order to consider the intermediate scattering problem (intermediate between Schrödinger and Klein-Gordon scattering) we need to derive a wave equation that unifies both the Klein-Gordon and Schrödinger equations. One approach to this is through the introduction of a fractional time derivative $\partial^q / \partial t^q, \;\; 1 < q < 2$ where $q = 1$ provides Schrödinger's equation and $q = 2$ yields the Klein-Gordon equation. A fractional partial differential equation that achieves this unification is (derived through induction)

$$\left(\nabla^2 - \frac{1}{c^q}\frac{\partial^q}{\partial t^q}\right) U = K_n U$$

where (c having fractional dimension $L^{2/q} s^{-1}$)

$$\frac{1}{c^q} = \left(\frac{2m}{i\hbar}\right)^{2-q} \frac{1}{c_0^{2(q-1)}}$$

and

$$K_n = \begin{cases} 2^{2-q}\kappa^2, & n = 1; \\ a^{2-q}(q-1)\kappa^{2(q-1)}, & n = 2. \end{cases}$$

The function K_n provides unification for the Schrödinger equation with $(n = 1)$ and without $(n = 2)$ the rest mass term, the constant a, with fractional dimension $L^{2(q-2)/(2-q)}$, being required to yield dimensional compatibility. With

$$\frac{1}{c^q} = \frac{1}{c_0^q}(1+\gamma)$$

we can then write

$$\left(\nabla^2 - \frac{1}{c_0^q}\frac{\partial^q}{\partial t^q}\right) U = \gamma \frac{1}{c_0^q}\frac{\partial^q U}{\partial t^q} + K_n U$$

where

$$\gamma = \left(\frac{2mc_0}{i\hbar}\right)^{2-q} - 1 = (-2i\kappa)^{2-q} - 1.$$

Defining a fractional differential in terms of the Fourier transform, i.e.

$$\frac{\partial^q}{\partial t^q} U(\mathbf{r}, t) \Longleftrightarrow (i\omega)^q u(\mathbf{r}, \omega),$$

we have

$$\left(\nabla^2 + \Omega^2\right) u = -\Omega^2 \gamma u + K_n u$$

where

$$\Omega^2 = -\frac{(i\omega)^q}{c_0^q}, \quad \Omega = \pm i\frac{(i\omega)^{q/2}}{c_0^{q/2}}.$$

The Born scattered field is then given by

$$u_s = \Omega^2 g(r,\omega) \otimes_3 \gamma u_i - g(r,\omega) \otimes_3 K_n u_i$$

where

$$g(r,\omega) = \frac{\exp(i\Omega r)}{4\pi r}.$$

The time dependent Green's function can be evaluated using the series expression for the complex exponential term by term as follows (taking $\Omega = -i(i\omega/c_0)^{q/2}$ to give consistency with the 'outgoing free space' Green's function in the case when $q = 2$):

$$G(r,t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} d\omega \exp(i\omega t)\frac{\exp[(i\omega/c_0)^{q/2} r]}{4\pi r}$$

$$= \frac{1}{4\pi r}\frac{1}{2\pi}\int_{-\infty}^{\infty} d\omega \exp(i\omega t)[1 + (i\omega/c_0)^{q/2} r$$

$$+\frac{1}{2!}(i\omega/c_0)^q r^2 + ...] = \frac{\delta(t)}{4\pi r} + \frac{1}{4\pi} c_0^{-q/2}\frac{\partial^{q/2}}{\partial t^{q/2}}\delta(t)$$

$$+\frac{1}{4\pi}\sum_{n=1}^{\infty}\frac{1}{(n+1)!} r^n c_0^{-(n+1)q/2}\frac{\partial^{(n+1)q/2}}{\partial t^{(n+1)q/2}}\delta(t).$$

Inverse Fourier transforming and using the convolution theorem, the time-dependent scattered field is given by

$$U_s = -\frac{1}{c_0^q}\frac{\partial^q}{\partial t^q}G(r,t) \otimes_3 \otimes_t \gamma U_i - G(r,t) \otimes_3 \otimes_t K_n U_i$$

where $\otimes_t$ denotes the convolution integral over t and U_s and U_i are the time-dependent scattered and incident fields respectively (i.e. the inverse Fourier transforms of u_s and u_i, respectively[5]). We note that for $r \to 0$,

$$U_s = -\frac{1}{4\pi r} \otimes_3 \gamma \frac{1}{c_0^q}\frac{\partial^q U_i}{\partial t^q} - \frac{1}{4\pi r} \otimes_3 K_n U_i$$

$$-\frac{1}{4\pi c_0^{3q/2}}\frac{\partial^{3q/2}}{\partial t^{3q/2}}\int_V \gamma(\mathbf{r}) U_i(\mathbf{r},t) d^3\mathbf{r}$$

$$-\frac{1}{4\pi c_0^{q/2}}\frac{\partial^{q/2}}{\partial t^{q/2}}\int_V K_n(\mathbf{r}) U_i(\mathbf{r},t) d^3\mathbf{r}$$

and that in the ultra-low frequency range (i.e. in the limit as $\omega_0 \to 0$),

$$u_s^0 = \frac{\Omega_0^2}{4\pi r} \otimes_3 \gamma - \frac{1}{4\pi r} \otimes_3 K_n.$$

In this case, the field U_s^0 is given by (for constant γ and κ)

$$U_s^0 = \frac{V}{4\pi r^2}(\Omega_0^2 \gamma - K_n)$$

which is zero when $\Omega_0^2 \gamma = K_n$ or when

$$k_0 = \frac{(-2)^{(q-2)/q} \kappa^{(q-2)/q} K_n^{1/q}}{\left(1 - \frac{(i/2)^{2-q}}{\kappa^{2-q}}\right)^{\frac{1}{q}}}.$$

9.13 Interpretation

If we define a gravitational field (for a spherically symmetric scatterer) to be given by the field U_s^0 then the interpretation of what gravity is must change. According to the universal scalar wavefield model considered here, a gravitational field is due to the scattering (by a material object composed of a spectrum of matter waves) of very low frequency scalar Helmholtz wavefields. Thus, if two bodies are in proximity, then each body will scatterer low frequency waves and each will interact with the scattered wavefield generated by the other, both experiencing an attractive (in the

[5] For notational convenience, we have used U_s to represent the time-dependent wavefield $U_s(\mathbf{r},t)$ which should not be confused with the use of $U_s(\mathbf{r},k), k \to 0$ in Section 9.7 or U_s^0 as used in Section 9.8.

non-relativistic case) gravitational force given by $v^2 m' U_s^0$ where m' is the mass of the other body. In this sense, we define gravity as follows:

Two bodies are attracted to each other because each 'detects' the 'gravity waves' scattered by the other in the non-relativistic case.

However, the term 'gravity waves' used here is not the same as that used in general relativity. The term relates to the low frequency components of a scalar wave spectrum and must be interpreted within the context of the limiting condition $k \to 0$.

The model provides results that are compatible with observable characteristics of a gravitational field: (i) a gravitational field is a weak field; (ii) a gravitational field is characterized by an inverse square law; (iii) a gravitational field deflects light; (iv) gravity is an attractive only force. However, in this model, the 'deflection' of light is not taken to be due to the bending of light as it travels through a curved space-time manifold (Einstein's model) but through the diffraction of light (and other electromagnetic radiation) by a gravitational field. It should be noted that, according to this model, gravity waves (as understood in terms of Einstein's equations) can not be measured. The attempt to detect Einstein gravity waves (i.e. the gravity waves predicted by general reativity) is the equivalent of constructing a weighing machine to weigh itself! Rather, we are 'detecting' gravity waves all the time, the effect of this 'detection' manifesting itself in terms of the 'force of gravity' we are all accustomed to.

The attractive only condition is valid for the non-relativistic case (i.e. for the Schrödinger scattering function). In the relativistic case, although the gravitational field U_s^0 is still weak, it depends on the square of the mass and generates a repulsive force. Note that in the case of the Schrödinger scattering function with potential energy E_p, then

$$\gamma > 0 \implies \frac{2mc_0^2(E - E_p)}{E^2} - 1 > 0$$

However, for any material characterised by a case when $E_p > E$, the scattering function is negative and the gravitational field defined by $\mathbf{U}_s^0$ will yield a repulsive force.

9.14 Principle of Eigenfield Tendency: Quantum Mechanics Revisited

Given the approach considered, an eigenfield tendency principle is required in order to explain the properties of matter as described by Schrödinger's equation (in the non-relativistic case) as originally conceived by Schrödinger [57]. For different potential energy functions $E_p(\mathbf{r})$, it is well known that this equation describes eigenfield systems that can be used to model the properties of matter through the principles of quantum mechanics (in the full context of the subject). The original reason for deriving the Schrödinger scattering function was so that the asymptotic behaviour of a

scattered Helmholtz wavefield (i.e. when $\omega \to 0$) could be examined. However, the consequence of this is that the Helmholtz equation is the governing wave equation only over a limited frequency band and that as the frequency of a wavefield increases (i.e. as $\omega \to \omega_1$) the Helmholtz equation reduces to the Schrödinger equation. If we consider the Schrödinger equation to represent eigenfields (at least in terms of its description of matter waves), then we can argue that at the higher end of the our universal spectrum, wavefields tend to behave more and more like eigenfields. Matter is thus taken to be composed of eigenfield systems at higher and higher frequencies; first the atom, then the nucleus, then the constituents of the nucleus (the quarks) and so on. Equations such as Schrödinger's equation and Dirac's equation are both descriptions for eigenfield systems at different energies (non-relativistic and relativistic energies respectively).

In the context of matter being an eigenfield system described by eigenfunction solutions to Schrödinger's equation, consider the case of a free electron and a free proton and the formation of hydrogen gas. In conventional (particle) terms, an electron and a proton have the same charge but of opposite polarity. This attracts the particles to form a neutral hydrogen atom, an effect which requires the introduction of a field, namely, an electric field. In terms of a wavefield theory, both the electron and proton are waves. In an ionised state, the electron is a free wave and the proton (relative to the electron) is a potential which is itself an eigenfield system (consisting of a higher frequency spectrum - the 'nuclear spectrum'). The free wavefield requires greater energy to exist in a free state and hence, based on the principle of least energy, will 'attempt to exist' as an eigenfield. This 'eigenfield' may have a number of eigenstates, each with a specific energy level. The difference in energy between the free energy state and the available eigenstate(s) provides a residual energy, i.e. a free energy wavefield with frequency $E/\hbar$. Once formed, the eigenfield will not share its eigenstate(s) as this will require greater energy and hence, if another electron comes in to the vicinity of the neutral hydrogen atom, it will appear to undergo a repulsive force. On the other hand, since the combined eigenfields associated with two hydrogen atoms requires lower energy than two separate eigenfields (i.e. two hydrogen atoms) then the result is the diatomic Hydrogen molecule H_2 - the result of a covalent bond. In this sense, an electric field is not the product of a charge, rather it is that entity associated with the propensity for a free wavefield to become an eigen wavefield. A magnetic field is then a measure of the rate of change over which this propensity is satisfied, i.e. If $U(\mathbf{r},t)$ exists such that

$$\int\int \mid U(\mathbf{r},t) \mid^2 d^3\mathbf{r}dt$$

is a minimum, then

$$\text{Free Wavefield} \quad \begin{matrix} \text{Electric Field } \mathbf{E} \\ \longrightarrow \\ \text{Magnetic field } \frac{\partial \mathbf{E}}{\partial t} \end{matrix} \quad \text{Eigen Wavefield}$$

Note that the transition described by Free Wavefield→ Eigen Wavefild may have both magnitude and direction since a free wavefield will attempt to find the shortest

possible path in a three-dimensional space in order to become an eigen wavefield. An electric field will therefore appear to be a vector field. Further, if the transition has no directional preference, then an electric field will appear to have a Coulomb field strength characterised by an inverse square law.

The principle of eigenfield tendency is just the principle of least energy as applied to a universal wavefield model. It is, however, a principle which allows us to explain an electric field without having to refer to the concept of a field being 'radiated' by a charge! For example, 'electron cloud' repulsion theory (Valence Shell Electron Pair Repulsion) is used to predict shapes and bond angles of simple molecules in which the 'electron cloud' may be a single, double or triple bond, or a lone pair of electrons - a non-bonding pair of electrons. The 'electron clouds' are taken to be negatively charged since the electrons are negatively charged, so electron clouds repel one another and try to get as far away from each other as possible. Instead of considering the electron cloud to consist of negatively charged electrons, we consider the cloud to be a eigenfield which arranges itself in such a way that it can exist in a minimum energy state, a state that affects the geometry of the molecule. In a simple hydrogen atom, for example, the eigenfield will be distributed symmetrically because, in a three-dimensional space, spherical symmetry represents the most energy efficient configuration which is equivalent to the electron wavefield 'experiencing' a Coulomb potential.

The eigenfunctions that are the solutions to the Schrödinger equation for different materials will not necessarily be complete eigenfunctions. In some cases, solutions only allow for the existence of quasi-eigenfunctions. In conventional atomic physics, quasi-eigenfunctions are incomplete standing waves more commonly referred to a delocalised electrons. These are electrons that exist in the 'lattice' of a material but are free to move and provides a material with the property we refer to as conductivity. This includes materials such as various metals and chemicals (e.g. Benzene which is composed of a ring of delocalised electrons). The principle difference between an eigenfield and a quasi-eigenfield, is that a quasi-eigenfield has an energy spectrum, albeit a narrow one.

The Schrödinger scattering function for matter waves is

$$\gamma = \frac{2mc_0^2(E - E_p)}{E^2} - 1.$$

In a macroscopic sense, E_p is the total potential energy associated with all the nuclei from which a material of compact support is composed and E is the total energy associated with the electrons. In the case of elements such as gold, the arrangement of electrons around the nucleus is such that a single electron occupies the outermost shell and is an example of a quasi-eigenfield, i.e. a relatively free wavefield (a free electron) that is only loosely bound to the host atom. Successive energy levels are contained in a small energy range dE and are so close that, in effect, a continuous energy spectrum is formed. Each energy level in this spectrum can accommodate a left-travelling and right-travelling wave ('spin-up' and 'spin-down' electrons - Pauli's principle) and these free electrons will distribute themselves throughout the energy band from 0 to some value E. Irrespective of any particular system, the number of

possible modes of oscillation per unit volume dn in a frequency range ν to $\nu + d\nu$ for waves with a propagation velocity of c is given by

$$dn = \frac{4\pi\nu^2 d\nu}{c^3}.$$

With $E = p^2/(2m) = \hbar\omega$ and $p = \hbar\omega/c = E/c$, then

$$dp = \frac{\hbar d\omega}{c} \text{ and } dE = \frac{p}{m}dp = \hbar d\omega.$$

The number of states per unit volume in the energy interval dE is therefore

$$dn(E) = \frac{(2m^3)^{\frac{1}{2}}E^{\frac{1}{2}}}{2\pi^2\hbar^3}dE$$

and thus, the total number of electrons per unit volume in the energy spectrum $(0, E)$ is[6]

$$n(E) = 2\frac{(2m^3)^{\frac{1}{2}}}{2\pi^2\hbar^3}\int_0^E E^{\frac{1}{2}}dE = 2\frac{(2m^3)^{\frac{1}{2}}}{3\pi^2\hbar^3}E^{\frac{3}{2}}.$$

Here m is taken to be the mass of an electron. Note that if the material is in a 'ground state' then the available electrons will occupy the lowest possible energy level. Further, if the total number of electrons per unit volume is less than the total number of energy levels available in a band (the bandwidth of the material), then the electrons can occupy all energy states up to a maximum energy E_{max} - the Fermi Energy. In this sense, the Fermi energy defines the (energy) bandwidth of a (conductive) material composed of a quasi-eigenfield.

With an atomic number of 79, gold is the heaviest of the most conductive elements in the periodic table, i.e. the product of the conductivity with the atomic number ($\sim 3.57 \times 10^7 \mathsf{cm}\Omega$) for gold is larger than any other element. If it were possible to reduce the total energy associated with the total quasi-eigenfield of gold such that $E < E_p$, then the result would be a scattering function that is negative. This requires the Fermi energy of gold to be reduced, the most influential factors being temperature and volume. Clearly, if the number of electrons per unit volume n is reduced then so is the Fermi energy. In terms of a physical material, n is determined by the number of atoms defining the physical extent of the material. This suggests an experimental investigation of the cryogenic properties of M-state (mono-atomic) gold. M-state gold is a white powder and is an example of a nano-material where each of the nano-metre size grains are clusters of a few hundred atoms. Like other M-state materials, the surface area is huge compared to the metallic (macro-crystalline) form. Thus, with the volume of each grain being small enough and the temperature of the material being low enough, it may be possibly to reduce the Fermi energy to an extent where $E < E_p$ for the material as a whole.

---

6 The factor of 2 is because of Pauli's principle.

9.15 Discussion

The results developed in this chapter encapsulate a phenomenology where the Helmholtz equation is, in effect, being used in an attempt to develop a unified scalar wavefield theory where the wavefield $u(\mathbf{r}, \omega)$ is taken to exist over a broad range of frequencies limited only by the Planck frequency. At very high frequencies, u is taken to describe matter waves which are characterised by relativistic (Klein-Gordon and Dirac equations) and non-relativistic energies (Schrödinger equation) associated with nuclear and atomic physics respectively. At intermediate frequencies, u is taken to describe waves in the 'electromagnetic spectrum' and at low frequencies, u is taken to describe waves in the 'gravity wave spectrum'.

The structure of matter, the characteristics of light and other electromagnetic radiation and the properties of gravity become phenomenologically related via Helmholtz scattering over different frequency bands. Low frequency waves (gravity generating waves) are scattered by high frequency waves (matter waves) to produce a gravitational field; intermediate frequency waves (electromagnetic spectrum) are scattered by high frequency waves (e.g. a lens) but can also be scattered by the field generated from the scattering of low frequency waves to produce gravitational diffraction. In this sense, 'physics' becomes the study of waves interacting with waves at vastly different frequencies, the breadth of the spectrum 'reflecting' the instantaneous birth of the universe - the 'big-bang' - since it requires (noting that the Fourier transform of a δ-function is a constant over all frequency space) a short impulse to generate a broad frequency spectrum. However, in attempting to derive a 'wavefield theory of everything' we must re-interpret the nature of an electric field using the principle of eigenfield tendency. Thus, instead of contemplating an electron in terms of a particle with a negative charge that 'radiates' an electric field and is attracted to particles with a positive charge (which also 'radiate' an electric field), we can visualise an electron in terms of a wave which is 'attracted' by the 'requirement' (through the minimum energy principle) of becoming an eigenfunction (a standing wave with lower energy than a free wave) whose properties are determined by the potential energy associated with the atomic nucleus which is itself, a higher (nuclear) frequency eigenfield system (quarks).

The form of the wave equation

$$\left(\nabla^2 - \frac{1}{c^2}\frac{\partial^2}{\partial t^2}\right) U(\mathbf{r}, t) = 0$$

dictates that c must be of finite value. If a wavefield (whatever the wavefield may be) was to convey information from one point in space to another instantaneously, then the second term of the above equation would be zero; the 'wave equation' would be reduced to 'Laplace's equation' $\nabla^2 u = 0$. Einstein's principal postulate is that the upper limit at which any wavefield can propagate is the speed of light c_0 in a perfect vacuum and thus $c \leq c_0$. In a more general perspective, the rationale associated with the fact that c must have a finite upper bound is that the influence of any physical wavefield on any measurable entity can only occur in a finite period

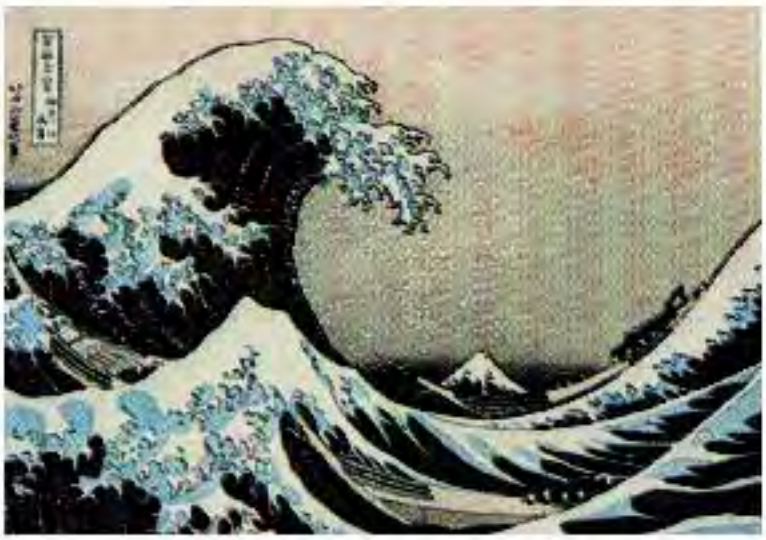

FIGURE 39 Example of fractal waves by the Japanese artist K Hokusai from the 1800s illustrating waves of different scale in both amplitude and wavelength.

of time and that there can be no such thing as instantaneous 'action at a distance', i.e. as Issac Newton put it: *That one body may act upon another at a distance through a vacuum, without the mediation of anything else, by and through which their action and force may be conveyed from one to the other, is to me so great an absurdity, that I believe no man who has in philosophical matters a competent faculty of thinking, can ever fall into it.* Taking Newton's own term, *mediation* requires the propagation (of information), but propagation at infinite speeds is not propagation and thus, we postulate that instantaneous fields are not possible, i.e. the speed at which a wavefield propagates must be finite for a wavefield to exist. In this context, the results developed for this thesis highlight the idea that the 'physics' of a wavefield is more fundamental than the 'physics' of a field. This principle should be considered in light of the fact that the one property common to the principal field equation of physics (e.g. Einstein's equations, Maxwell's equations, Proca's equations), is that they all describe wave phenomena - at least in an 'indirect' sense. In the case of Proca's equations, the field equations are derived with the singular aim of ensuring that they can be decoupled to yield the inhomogeneous Klein-Gordon (wave) equation.

9.15.1 Fractal Wave Model

The underlying philosophy associated with the approach considered, is based on a 'waves within waves' model, i.e. to quote an old Chinese proverb *'In every way, one can see the shape of the sea'.* This is a universal self-affine or fractal model in which the 'fractal field' is a scalar wavefield, a symbolic representation of the idea being given in Figure 39. As the frequency increases, a wavefield tends to become an eigenfield. This principle is required to explain the structure of matter and much of the discussion given in Section 9.13 is quantum mechanics revisited without the need

to define an electric field in terms of a charge. If we consider the structure of matter at the atomic, nuclear and sub-nuclear scales (indeed at all scales down to the scale of the Planck length) to be determined by eigenfields, then the question remains as to why eigenfield systems should 'kick-in' at the atomic scale? If the principle of eigenfield tendency applies at all frequencies then why do we not observe equivalent naturally occurring eigenfield systems in the electromagnetic spectrum? Perhaps we do under special circumstances, e.g. ball-lightning.

The approach to unification considered has yielded a number of questionable and speculative results. The only experimental evidence offered in confirmation to our model for a gravitational field is a possible explanation as to why the Einstein rings associated with near field galaxies observed by the Hubble Space Telescope are blue. However, it should be noted that this 'evidence' is most typical of Carl Popper's principle that all observation statements are 'theory laden' and that other explanations may be possible that are more appropriate in terms of established physical models.

In general relativity, the curvature of space-time bends light by the same amount irrespective of the frequency - there is no dispersion relation. The λ^{-6} scaling law associated with gravitational diffraction may be validated (or otherwise) from appropriate simultaneous observations of the same Einstein ring (complete or otherwise) at different wavelengths. Other consequences such as a gravitational field generating a repulsive force that is proportional to the mass squared in the relativistic case remain of theoretical consequence only. However, it is noted that inflation theory (the expansion of the early universe) requires gravity to be a repulsive force.

The model considered leads to the proposition that a gravity field is regenerative and exists through the continuous scattering of existing low frequency Helmholtz wavefields. This proposition may provide an answer to the following question: If nothing can escape the event horizon of a black hole because nothing can propagate faster than light then how does gravity get out of a black hole? The conventional answer to this question is that the field around a black hole is 'frozen' into the surrounding space-time prior to the collapse of the parent star behind the event horizon and remains in that state ever after. This implies that there is no need for continual regeneration of the external field by causal agents. In other words, the explanation defies causality. In the model presented here, the gravitational field generated by a black hole or any other body is the result of a causal effect - the scattering of low frequency scalar waves. In this sense, a black hole is just a stronger scatterer than other cosmological bodies and a gravitational field 'gets out of a black hole' because it was never 'in the black hole' to start with.

9.15.2 Propagative Theories

Propagative or wave theories of gravity have been proposed for many years. In 1805, Laplace proposed that gravity is a propagative effect and considered a correction to Newton's law to take into account the observation that gravity has no detectable aberration or propagation delay for its action. Laplace's ideas were advanced further by Weber, Riemann, Gauss and Maxwell in the Nineteenth Century using a variety

of 'corrective terms'. In 1898, Gerber, developed a propagative theory that took into account the perihelion advance of mercury and in 1906 Poincaré showed that the Lorentz transform cancels out gravitational aberration. After the success of general relativity (1916) for explaining gravity in terms of a geometric effect, propagation theories were discarded. However, more recently, attempts at explaining gravity in terms of causal effects through a 'propagative' force have been revisited [81] as debate over the basic Einsteinian postulates[7] has intensified. Moreover, from Laplace to the present, propagation theories of gravity consider an object to be 'radiating' a field (in a passive sense). If general relativity considers gravity to be the result of an object warping space-time, then the proposition reported is that gravity is the result of an object scattering (long wavelength) waves that already exist as part of the low frequency component of a universal spectrum which is, itself, the by-product of the 'big-bang'.

9.15.3 Compatibility with General Relativity

The compatibility of this approach with general relativity can be realised if the wavefield as taken to warp space-time so that space-time is the medium of propagation. Only at very large wavelengths does the warping of space-time become so pronounced and over such a large scale that Einstein's field equations can then be used to describe the physics associated with the geometry of the field. In other words, if space-time is taken to be the medium of propagation of all (scalar) wavefields at all frequencies, then the theory of general relativity emerges naturally as $k \rightarrow 0$. A two-dimensional and qualitative illustration of this idea is given in Figure 40 which shows four frames of a simple two-dimensional wave function as $k \rightarrow 0$. It is assumed that the wavefunction is due to the scattering of a plane wave from a delta function located at the centre of the surface. If space is taken to be the medium of propagation which undergoes curvature as a wave propagates through it then Figure 40 can be taken to illustrate the curvature of a two-dimensional space into a three dimensional space at increasingly lower frequencies. As $k \rightarrow 0$ the wavefield is replaced by what appears to be a static curved space manifold within the locality of a low frequency scattering event. The curvature of this manifold is the taken to be responsible for generating a gravitation force which is attractive in terms of the influence of one mass upon another and is compounded in terms of Einstein's field equation, i.e.

$$R_{\mu\nu} - \frac{1}{2} g_{\mu\nu} R + g_{\mu\nu} \Lambda = \frac{8\pi}{c^4} T_{\mu\nu}$$

where $R_{\mu\nu}$ is the Ricci curvature tensor, R is the scalar curvature, $g_{\mu\nu}$ is the metric tensor, Λ is the consmological constant, G is the gravitational constant, c is the speed of light and $T_{\mu\nu}$ is the stress-energy tensor [56].

Any propagation theory of gravity must address some basic known observations:

---

[7] The invariance of the propagation of light in a vacuum for any observer which amounts to a presumed absence of any preferred reference frame.

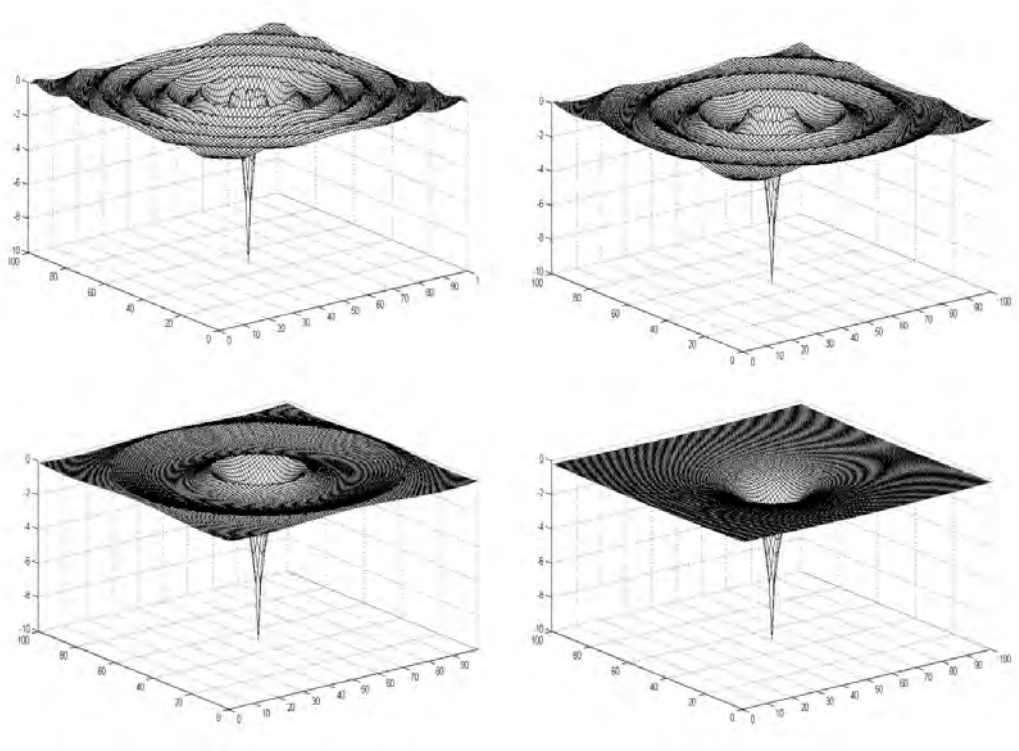

FIGURE 40 Qualitative illustration of the function $-\mathrm{Re}[\cos(kr)/r]$, $r = \sqrt{x^2 + y^2}$ for four frames as $k \to 0$ (from left to right and from top to bottom).

- Gravity has no detectable aberration or propagation delay for its action leading to effects predicted by general relativity such a *gravitomagnetism*;
- the finite propagation of light causes radiation pressure for which gravity has no counterpart pressure.

These results represent the most vital evidence with regard to gravity being a geometric and not a propagative effect. For example, in an eclipse of the Sun, the gravitational pull on the earth by this 3-body (Sun-Moon-Earth) configuration increases. By comparing the delay in time it takes to observe the visible maximum eclipse on Earth (which can be calculated from knowledge of the distance of the Moon from the Earth) with the equivalent gravitational maximum, then if gravity is a propagating force, it appears to propagates at least 20 times faster than light! [82] Irrespective of whether this value is valid or not, a fundamental issue remains, which is compounded in the question: what is the speed of gravity? If we consider gravity to be a propagation and/or a low frequency scattering effect, then in order to account for the lack of propagation delay, it must be assumed that the speed of gravity is greater than the speed of light. This is contrary to the Einsteinian postulates if these postulates are taken to apply to all wavefields irrespective of their wavelength. The model presented here assumes that the speed of gravity is the same as the speed of light c_0. However, the asymptotic result $k \to 0$ used to define a gravitational field yields, what will appears to be, an instantaneous effect from a wavefield that is taken to propagate at the speed of light. The wavelength is so long

compared to the distances associated with a Sun-Moon-Earth system, for example, that the speed of gravity will appear to be significantly faster than the speed of light (i.e. $\mathbf{U}_s^0$ is observed to be an instantaneous field).

9.16 Final Comments

In terms of the fractal wavefield model considered here, the gravitational force is a consequence of very long wavelength waves and is therefore a long range force. Electromagnetism is a consequence of intermediate wavelength waves which exist as both free wavefields and eigen wavefields at the atomic scale, the transition from one to the other creating an 'electric field'. The strong force is a consequence of a nuclear eigen wavefield where the values of $E = \hbar\omega$ and $p = \hbar k$ are in the relativistic energy limit. The weak force (associated with radioactive decay, for example) is explained in terms of the transformation of a nuclear eigen wavefield to a more stable form allowing for the emission of a free wavefield (quantum 'tunneling effect' when the potential barrier is low). For example, Rutherford scattering (the scattering of alpha particles from gold nuclei which historically provided the basic model for the atom) is an example of a free (nuclear) wavefield, interacting with a stable eigenfield system which consequently appears to exert a repulsive Coulomb force. At this frequency range the governing equation is Schrödinger's equation which has a far field scattering amplitude determined by the three-dimensional Fourier transform of a Coulomb potential. Thus, as a function of the scattering angle θ

$$A(\theta) = \frac{2\pi}{k \sin\left(\frac{\theta}{2}\right)} \int_0^\infty \sin\left[2kr \sin\left(\frac{\theta}{2}\right)\right] \gamma(r) r dr$$

and for the screened Coulomb potential[8]

$$\gamma(r) = \frac{\exp(-ar)}{r}, \quad a > 0$$

we obtain (for $a \to 0$)

$$A(\theta) = \frac{\pi}{k^2 \sin^2\left(\frac{\theta}{2}\right)} \left(1 + \frac{a^2}{\left[2k \sin\left(\frac{\theta}{2}\right)\right]^2}\right)^{-1} = \frac{\pi}{k^2 \sin^2\left(\frac{\theta}{2}\right)}.$$

The intensity (scattering cross-section) is therefore inversely proportional to $\sin^4(\theta/2)$ which is the basic 'signature' of Rutherford scattering. In terms of neutron scattering, a neutron is a free nuclear wavefield which, during its life time, is unable to combine with an existing nuclear eigen wavefield until it does, in some cases producing unstable nuclear eigen wavefield systems which transform into new stable systems involving the emission of free wavefields, i.e. nuclear fission.

[8] Required in order evaluate the integral over r.

Note that the principle of eigenfield tendency in which free wavefields tend to become eigen wavefield in order to achieve a minimum energy is equivalent to the least action principle. In field theory - in this case, the wavefield $U(\mathbf{r},t)$ - the Lagrangian density $\mathcal{L}$ is a functional that is integrated over all space-time, i.e.

$$\mathcal{S}[U] = \int\int \mathcal{L}[U, \partial_\mu U] d^3\mathbf{r} dt$$

where, using 'relativistic notation',

$$\partial_\mu = (\partial^0; \nabla), \quad \partial^\mu = (\partial^0; -\nabla),$$

$$\partial^0 = \frac{1}{c}\frac{\partial}{\partial t} \quad \text{and} \quad \partial_\mu \partial^\mu = \frac{1}{c^2}\frac{\partial^2}{\partial t^2} - \nabla^2.$$

The Lagrangian is the spatial integral of the density and application of the least action principle yields the Euler-Lagrange equations

$$\frac{\delta \mathcal{S}}{\delta U} = -\partial_\mu \left(\frac{\partial \mathcal{L}}{\partial(\partial_\mu U)} \right) + \frac{\partial \mathcal{L}}{\partial U} = 0$$

which are then solved for U.

The wavefield approach adopted is consistent with the basic concepts associated with the *Grand Unified Theories* of C H Tejman [83] and in one sense, we have attempted to explain the example images given in Figure 36 using a single phenomenological model. Just as Poisson used a wave model to explain the Poisson spot without reference to light being an electromagnetic wave (Maxwell's equations for an electric and magnetic field which Poisson did not know of at the time), so we have attempted to explain both a Poisson spot and an Einstein ring without reference to general relativity (Einstein's equation for a gravitational field). The problem then remains of how to 'formally recover' Maxwell's equations and Einstein's equations from a single wave theoretic model.

10 DISCUSSION AND CONCLUSIONS

10.1 Discussion

The principal theme of this thesis has been to explore the EM scattering problem in an attempt to develop models that incorporate strong scattering for which the inverse scattering problem becomes a feasible proposition. Four approaches have been considered in this respect:

- implementation of the exact inverse scattering solutions considered in Chapter 6;
- application of diffusion models as discussed in Chapter 7;
- application of fractional diffusion to model the intermediate case as given in Chapter 8;
- low frequency scattering as considered in Chapter 9.

The exact inverse scattering theory considered is based on modifying the inhomogeneous Helmholtz equation to the form

$$-k^2\gamma(\mathbf{r}) = \frac{u^*(\mathbf{r},k)}{\mid u(\mathbf{r},k) \mid^2}\nabla^2\left(u_s(\mathbf{r},k) - \frac{k^2}{4\pi r}\otimes_3 u_s(\mathbf{r},k)\right)$$

For far field applications (the most typical case), given that

$$\|u_s - (k^2/4\pi r)\otimes_3 u_s\|_2 \leq \|u_s\|_2[1 + k^2\sqrt{r/(4\pi)}],$$

we have considered the result

$$-k^2\gamma = \frac{-1}{u_i^{\pm} + u_s}k^2 u_s \otimes_3 \nabla^2\left(\frac{1}{4\pi r}\right) = k^2 \mid u_i^{\pm} + u_s \mid^{-2} [(u_i^{\pm})^* + u_s^*]u_s, \;\; r \to \infty$$

Working in one-dimension, this result has led to a model for the signal $s(t)$ generated by a pulse-echo system (with Impulse Response Function $p(t)$ and carrier frequency ω_0) of the form

$$s(t) = p(t) \otimes [\epsilon_r(t) - 1]\exp(-i\omega_0 t) + p(t) \otimes [\mid s(t) \mid^2 \exp(-i\omega_0 t)]$$

where $\epsilon_r(t)$ is the inhomogeneous permittivity profile as a function of the two-way travel time t. The second term in the above expression is taken to be the component due to multiple scattering processes which contributes to the noise term in the conventional model for a signal under the weak scattering approximation. However, this result relies on the condition that the band-width of the Impulse Response Function is significantly small compared to the carrier frequency and thus, the result conforms to side-band systems only. The result is also based on the assumption that $\mid u_i^{\pm} + u_s \mid^{-2} \sim 1$ and in order to compute the signal, it is necessary to iterate. However, the inverse scattering solution is not iterative and can be applied directly to evaluate the permittivity profile $\gamma(t)$ given $s(t)$.

Application of the diffusion based models developed in Chapter 7 is accomplished using the following transformation from the wave equation to the diffusion equation:

$$\left(\nabla^2 - \frac{1}{c^2}\frac{\partial^2}{\partial t^2}\right) u(\mathbf{r},t) = 0, \quad u(\mathbf{r},t) = \phi(\mathbf{r},t)\exp(i\omega t)$$

$$\downarrow$$

$$\left(\nabla^2 - \frac{1}{D}\frac{\partial}{\partial t}\right) \mid \phi(\mathbf{r}t) \mid^2 = 0$$

under the conditions that

$$\left|\frac{\partial^2\phi}{\partial t^2}\right| << 2\omega \left|\frac{\partial\phi}{\partial t}\right| \quad \text{and} \quad \mathrm{Re}[\nabla\cdot(\phi\nabla\phi^*)] = 0$$

The principal condition associated with this transformation is that the field ϕ varies significantly slowly in time compared with $\exp(i\omega t)$.

Inverse solutions to the problem 'Given $\mid \phi(\mathbf{r}, t = T) \mid^2$ compute $\mid \phi(\mathbf{r}, t = 0) \mid^2$' have been developed which are compounded in the implementation of well defined Finite Impulse Response filters of different order according to a Taylor series expansion of $\mid \phi(\mathbf{r}, t = T) \mid^2$. This diffusion based approach provides a method of solving the inverse scattering problem for strong scattering interactions. This is because the diffusion equation is based on a random walk process in which the scattering angle is uniformly distributed as in the calculation of the K-distribution given in Section 7.2.

In Chapter 8, the problem is modelling intermediate scattering has been considered based on the following transformation from the diffusion to the fractional diffusion equation:

$$\left(\nabla^2 - \frac{1}{D}\frac{\partial}{\partial t}\right) \mid \phi(\mathbf{r},t) \mid^2 = 0$$

$$\downarrow$$

$$\left(\nabla^2 - \frac{1}{D^q}\frac{\partial^q}{\partial t^q}\right) \mid \phi(\mathbf{r},t) \mid^2 = 0$$

where $1 < q < 2$. The use of fractional calculus has then be explored to obtain a Green's function for this equation. This is an entirely phenomenological approach based on the concept of a random walk with a directional bias as presented in Sections 8.1 and 8.2.

The ideas developed in Chapter 9 were originally based on considering low frequency scattering theory given that the Born series solution to the inhomogeneous Helmholtz equation can be reduced to a single and exact scattering transform for the case when $k \to 0$. However, although this asymptotic solution provides an exact scattering transform (and thereby an exact inverse scattering transform), it is not of any practical significance in imaging systems engineering. Instead, Chapter 9 develops a hypothesis which extends beyond the original scope of this thesis and considers the foundations of an approach to developing a unified wavefield theory based on the premise that all physical forces are manifestations of wavefields interacting with other wavefields over a broad frequency spectrum. This hypothesis is based on a philosophical extension of the idea that lies behind the Maxwell-Proca equations in which Proca introduces new terms into Maxwell's equations in order that they decouple to produce a relativistic wave equation (the Klein-Gordon equation) rather than a non-relativistic wave equation.

10.2 Conclusions

Of the material presented in this work, the most innovative is the use of the fractional diffusion equation to model intermediate scattering events. In terms of the theory of scattering from deterministic media and compatibility with the imaging equation, the Born approximation represents a central theme. The exact inverse scattering theory developed for this thesis (in Chapter 4) and investigated further in Chapter 6 provides the potential for improving image reconstruction and image processing methods in general. For example, a principal result of the material presented in Chapter 6 is that speckle reduction is in coherent pulse-echo imaging systems should be applied to the complex data before the amplitude image is computed in order to reduce the cross terms associated with single and multiple scattering processes.

With regard to the EM scattering from random media, three approaches have been reviewed:

- application of weak scattering theory to a random scatterer which can be cast in terms of computing the Fourier transform of a cross-correlation function model for the random scatterer;
- application of statistical modelling methods for computing the PDF of the scattered intensity based on random walk models applied to the amplitude and phase;
- application of the diffusion equation

The use of the weak scattering approximation for scattering from random media suffers from limitations in that multiple scattering is assumed to be negligible. Direct statistical modelling is therefore preferable as it can take into account the case of multiple scattering processes. However, while this approach may provide a valid model for the PDF of a signal or image that can be used for statistical image

analysis, for example, it is not of value in developing inverse scattering solutions that are of value to processing a signal or image for the retrieval of information. On the other hand, application of diffusion models for strong scattering leads to inverse solutions (FIR filters) that can be used to reconstruct an image. For this reason, the fractional diffusion equation has been used to develop inverse solutions to the generalised problem and one of the principal conclusions of this thesis is that fractional diffusion models provide a useful generalised approach to modelling EM scattering problems that has practical value in the processing and interpretation of EM signals and images.

10.3 Open Problems

1. Development of an intermediate scattering theory based on taking the Fresnel transform of equation (4.8)

$$\gamma = A^{-1}[(u_i^{\pm})^* + u_s^*]u_s, A^{-1} = \mid u_i^{\pm}) + u_s \mid^{-2}$$

   under application of the condition $\widetilde{A}^{-1}$ and the Skew Hermitian condition considered in Sections 4.8.1 and 4.8.2 using the Fourier transform.

2. Development of a near-field scattering theory using equation (4.7)

$$-k^2\gamma(\mathbf{r}) = \frac{u^*(\mathbf{r},k)}{\mid u(\mathbf{r},k) \mid^2}\nabla^2\left(u_s(\mathbf{r},k) - \frac{k^2}{4\pi r} \otimes_3 u_s(\mathbf{r},k)\right)$$

   based on application of the convolution transform

$$\gamma(\mathbf{r}) \otimes_3 g(r,k)u_i^{\pm}(\mathbf{r},k)$$

   for the computation of u_s.

3. The simulations undertaken in Section 4.8.3 have been based on considering the first order iteration to compare the scattered field under the Born approximation with the effect of multiple scattering defined in terms of the autoconvolution of a Born scattered field. The effect of undertaking further iterations should be investigated and the relationship of each iteration with the physical nature of the scattered field quantified and compared with the interpretation of Born series given in Section 4.5.1.

4. Computation of the scattered field using the complex scattering function $-k^2\gamma + ikz_0\sigma$.

5. Application of the exact inverse scattering solutions developed in Chapter 4 (based on the material given in Appendix 1) for designing image reconstruction algorithms used in diffraction tomography and other optical and electromagnetic imaging system.

6. Coherent signal extraction based on the strong scattering model developed in Chapter 6, i.e.
$$s(t) = p(t) \otimes [\gamma(t) \exp(-i\omega_0 t)] + n(t)$$
where
$$n(t) = p(t) \otimes [\mid s(t) \mid^2 \exp(-i\omega_0 t)]$$
and a systematic analysis of the performance for de-noising coherent signals based on this model.

7. Extension of the weak scattering SAR model developed in Chapter 5 to include strong scattering effects based on the model developed in Chapter 6.

8. Extension of random Born scattering model considered in Section 7.1 to include multiple scattering effects based on equation (4.8) and comparison of the statistical properties of the field with K-scattering discussed in Section 7.2.

9. Derivation of diffusion based models (as discussed in Section 7.3) for coherent scattering processes and associated inverse solutions.

10. Application of fractional diffusion models for processing coherent signals and images generated by the scattering of coherent radiation from random media for intermediate strength scattering processes.

11. Derive fractional diffusive models that are based on more fundamental principles and associated derivations than the phenomenological arguments presented in Section 8.2

12. Experimental verification or otherwise of the λ^{-6} scaling law for the diffraction of electromagnetic waves by a gravitational field.

REFERENCES

[1] W. Coffey, Y. P. Kalmykov and J. T. Waldron, *The Langevin Equation: With Applications in Physics, Chemistry, and Electrical Engineering*, World Scientific, 1996.

[2] J. C. Maxwell, *A Dynamical Theory of the Electromagnetic Field*, Philosophical Transactions of the Royal Society of London **155**, 459-512, 1865.

[3] P. M. Morse and H. Feshbach, *Methods of Theoretical Physics*, McGraw-Hill, 1953.

[4] J. A. Stratton, *Electromagnetic Theory*, McGraw-Hill, 1941.

[5] M. Fujimoto, *Physics of Classical Electromagnetism*, Springer, 2007.

[6] S. Cloude, *An Introduction to Electromagnetic Wave Propagation and Antennas*, Springer, 1996.

[7] H. G. Green, *A Biography of George Green*, in *Studies and Essays in the History of Science and Learning*, Arno Press, 1975.

[8] N. M. Ferrers, *Mathematical Papers of George Green*, Chelsea, 1970.

[9] D. M. Cannell, *George Green: Mathematician and Physicist 1793-1841: The Background to His Life and Work*, Society for Industrial and Applied Mathematics, Philadelphia, 1993.

[10] P. M. Morse and H. Feshbach, *Methods of Theoretical Physics*, 1953, McGraw-Hill.

[11] G. F. Roach, *Green's Functions (Introductory Theory with Applications*, Van Nostrand Reihold, 1970.

[12] E. N. Economou, *Green's Functions in Quantum Physics*, Springer-Verlag, 1979.

[13] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*, National Bureau of Standards Applied mathemtics Series 55, 1964
http://www.iopb.res.in/ somen/abramowitz_and_stegun

[14] R, G. Newton, *Inverse Schrödinger Scattering in Three Dimensions*, Springer, 1989.

[15] P. M. Morse and H. Feshbach H, *Methods of Theoretical Physics*, McGraw-Hill, 1953.

[16] R. G. Newton, *Scattering Theory of Waves and Particles*, Springer-Verlag, 1967.

[17] E. Butkov, *Mathematical Physics*, Addison-Wesley, 1973.

[18] J. M. Blackledge, *Digital Image Processing*, Horwood Publishing, 2005 (Chapter 11).

[19] A. W. Rihaczek, *Principles of High Resolution Radar*, McGraw-Hill, 1969.

[20] R. O, Harger, *Synthetic Aperture Radar Systems*, Academic Press, 1970.

[21] J. J. Kovaly, *Synthetic Aperture Radar*, Artech, 1976.

[22] R. L. Mitchell, *Radar Signal Simulation*; MARK Resources, 1985.

[23] J. M. Blackledge, *Digital Signal Processing* Second Edition, Horwood Publications, 2006.

[24] P. A. Martin, *Multiple Scattering: Interaction of Time-Harmonic Waves with N Obstacles*, Encyclopedia of Mathematics and its Applications 107, Cambridge University Press, 2006.

[25] The MathWorks, *Digital Signal Processing*, 2009.
http://www.mathworks.com/applications/dsp_comm/

[26] D. J. Griffiths, *Introduction to Quantum Mechanics* (Second Edition), Prentice Hall, 2004.

[27] S. Winitzki, *Cosmological Particle Production and the Precision of the WKB Approximation*, Physical Review D 72: 104011, 2005

[28] R. Jost and W. Kohn, *Construction of a Potential from a Phase Shift*, Phys. Rev. **37**, 977-992, 1952.

[29] R. T. Prosser, *Formal Solutions of Inverse Scattering Problems*, J. Math. Phys. **17**, 1175-1779, 1976

[30] Sandia National Laboratories, *Sandia Complex Image Data*, Synthetic Aperture Radar, 2009; http://www.sandia.gov/RADAR/sar-data.html

[31] M. Bertero and B. Boccacci, *Introduction to Inverse Problems in Imaging*, Institute of Physics Publishing, 1998.

[32] R. H. T. Bates and M. J. McDonnal, *Image Restoration and Reconstruction*, Oxford Science Publications, 1986.

[33] R. C. Gonzalez and P. Wintz, *Digital Image Processing*, Addison-Wesley, 1987.

[34] M. J. Turner, J. M. Blackledge and P. Andrews, *Fractal Geometry in Digital Imaging*, Academic Press, 1997.

[35] P. M. Morse and H. Feshbach, *Methods of Theoretical Physics*, McGraw-Hill, 1953.

[36] A. Einstein, *On the Motion of Small Particles Suspended in Liquids at Rest Required by the Molecular-Kinetic Theory of Heat*, Annalen der Physik, Vol. 17, 549-560, 1905.

[37] A. Einstein, *Investigations On The Theory Of The Brownian Movement*, Dover, 1956.

[38] J. M. Blackledge, G. A. Evans and P. Yardley, *Analytical Solutions to Partial Differential Equations*, Springer, 1999.

[39] H. Hurst, *Long-term Storage Capacity of Reservoirs*, Transactions of American Society of Civil Engineers, Vol. 116, 770-808, 1951.

[40] M. F. Shlesinger, G. M. Zaslavsky and U. Frisch (Eds.), *Lévy Flights and Related Topics in Physics*, Springer 1994.

[41] R. Hilfer, *Foundations of Fractional Dynamics*, Fractals Vol. 3(3), 549-556, 1995.

[42] A. Compte, *Stochastic Foundations of Fractional Dynamics*, Phys. Rev E, Vol. 53(4), 4191-4193, 1996.

[43] T. F. Nonnenmacher, *Fractional Integral and Differential Equations for a Class of Lévy-type Probability Densities*, J. Phys. A: Math. Gen. Vol. 23, L697S-L700S, 1990.

[44] R. Hilfer, *Exact Solutions for a Class of Fractal Time Random Walks*, Fractals, Vol. 3(1), 211-216, 1995.

[45] R. Hilfer and L. Anton, *Fractional Master Equations and Fractal Time Random Walks*, Phys. Rev. E, Vol. 51(2), R848-R851, 1995.

[46] S. Mallat, *A Wavelet Tour of Signal Processing*, Academic Press, ISBN: 0-12-466606-X, 1999.

[47] R. H. T. Bates and M. J. McDonnal, *Image Restoration and Reconstruction*, Oxford Science Publications, 1986.

[48] M. Bertero and B. Boccacci, *Introduction to Inverse Problems in Imaging*, Institute of Physics Publishing, 1998.

[49] M. J. Turner, J. M. Blackledge and P. Andrews, *Fractal Geometry in Digital Imaging*, Academic Press, 1997.

[50] B. B. Mandelbrot, *The Fractal Geometry of Nature*, Freeman, 1983.

[51] K. J. Falconer, *Fractal Geometry*, Wiley, 1990.

[52] L. Smolin *The Trouble with Physics: The Rise of String Theory, the Fall of a Science, and What Comes Next*, Houghton-Mifflin, 2006, ISBN-10: 0-618-55105-0 (http://www.thetroublewithphysics.com/).

[53] P. Woit *Not Even Wrong: The Failure of String Theory and the Search for Unity in Physical Law*, Basic Books, 2006, ISBN: 0-465-09275-6.

[54] B. B. Mandelbrot, *The Fractal Geometry of Nature*, Freeman, 1983.

[55] J. C. Maxwell, *A Dynamical Theory of the Electromagnetic Field*, Philosophical Transactions of the Royal Society of London **155**, 459-512, 1865.

[56] A. Einstein, *The Foundation of the General Theory of Relativity*, Annalen der Physik, IV, Floge 49, 770-822, 1916,
http://www.alberteinstein.info/gallery/gtext3.html.

[57] E. Schrödinger, *Quantization as an Eigenvalue Problem*, Annalen der Physik, **489**, 79, 1926.

[58] P. A. M. Dirac, *The quantum theory of the electron* Proc. R. Soc. (London) A, **117**, 610-612, 1928.

[59] P. A. M. Dirac, *The quantum theory of the electron: Part II* Proc. R. Soc. (London) A, **118**, 351-361, 1928.

[60] P. M. Morse and H. Feshbach, *Methods of Theoretical Physics*, McGraw-Hill, 1953.

[61] J. J. Sakurai, *Advanced Quantum Mechanics*, Addison Wesley, 1967, ISBN: 0-201-06710-2.

[62] A. S. Davydov, *Quantum Mechanics (2nd Edition)*, Pergamon, 1976, ISBN: 0-08-020437-6.

[63] W. Rarita and J. Schwinger, *On a Theory of Particles with Half-Integral Spin*, Phys. Rev. 60, 61-61, 1941.

[64] A. Proca, *Fundamental Equations of Elementary Particles*, C. R. Acad. Sci. Paris, 202, 1490, 1936.

[65] W. Greiner, *Relativistic Quantum Mechanics (3rd edition)*, Springer, 2000.

[66] J. A. Stratton, *Electromagnetic Theory*, McGraw-Hill, 1941.

[67] R. H. Atkin, *Theoretical Electromagnetism*, Heinemann, 1962.

[68] D. J. Griffiths, *Introduction to Quantum Mechanics (Second Edition)*, Prentice Hall, 2004 .

[69] R. Tomaschitz, *Einstein Coefficients and Equilibrium Formalism for Tachyon Radiation*, Physica A, **293**, 247-272, 2001.

[70] R. Tomaschitz, *Tachyon Synchrotron Radiation*, Physica A, **335**, 577-610, 2004.

[71] R. Tomaschitz, *Quantum Tachyons*, Eur. Phys. J. **D32**, 241-255, 2005.

[72] X. Bei, C. Shi and Z. Liu, *Proca Effect in Kerr-Newman Metric*, Int. J. of Theoretical Physics, **43**, 1555-1560, 2004.

[73] R. Scipioni, *Isomorphism Between Non-Riemannian Gravity and Einstein-Proca-Wyle Theories Extended to a Class of Scalar Gravity Theories*, Class. Quantum Gravity, **16**, 2471-2478, 1999.

[74] J. M. Blackledge, *Digital Image Processing*, Horwood Publishing, 2006, ISBN: 1-898563-49-7.

[75] G. F. Roach, *Green's Functions (Introductory Theory with Applications)*, Van Nostrand Reihold, 1970.

[76] Wikipedia, *Arago Spot*, 2009.
http://en.wikipedia.org/wiki/Arago_spot

[77] Wikipedia, *Einstein Ring*, 2009.
http://en.wikipedia.org/wiki/Einstein_ring

[78] F. Cain, *Near Perfect Einstein Ring Discovered*, Universe Today, 2005.
http://www.universetoday.com/2005/04/29/near-perfect-einstein-ring-discovered/

[79] V. Belokurov et al., *The Cosmic Horseshoe: Discovery of an Eistein Ring Around a Giant Luminous Red Galaxy*, Astrophysical Journal (Submitted), 2007 (http://www.arxiv.org/abs/0706.2326).

[80] A. Bolton, *Astronomy Picture of the Day*, 2008.
http://apod.nasa.gov/apod/ap080728.html

[81] T. C. Van Flandern, *Dark Matter, Missing Planets and New Comets: Paradoxes Resolved, Origins Illuminated*, North Atlantic Books, Berkeley, 1993

[82] T. C. Van Flandern, *The Speed of Gravity: What the Experiments Say*, Physics Letters A, 250, 1-11, 1998.

[83] C. H. Tejman, *Grand Unified Theory*, 2009.
http://www.grandunifiedtheory.org.il/

[84] K. B. Oldham and J. Spanier, *The Fractional Calculus*, Academic Press, 1974.

[85] A. Dold and B. Eckmann (Eds.), 1975, *Fractional Calculus and its Applications*, Springer, 1975.

[86] K. S. Miller and B. Ross, *An Introduction to the Fractional Calculus and Fractional Differential Equations*, Wiley, 1993.

[87] S. G. Samko, A. Kilbas and O. I. Marichev, *Fractional Integrals and Derivatives: Theory and Applications*, Gordon and Breach, 1993.

[88] V. Kiryakova, *Generalized Fractional Calculus and Applications*, Longman, 1994.

[89] I. N. Sneddon, *The use in Mathematical Physics of Erdélyi-Kober operators and of some of their Generalizations*, Lectures Notes in Mathematics (Eds. A Dold and B Eckmann), Springer, 37-79, 1975.

APPENDIX 1 EXACT INVERSE SCATTERING SOLUTIONS

APPENDIX 1.1 Exact Inverse Scattering Solution in One-Dimension

Consider the 1D inhomogeneous Helmholtz equation for a scalar (complex) wavefield $u(x,k)$ given by

$$\left(\frac{\partial^2}{\partial x^2}+k^2\right)u(x,k) = -k^2\gamma(x)u(x,k), \quad x \in (-\infty,\infty)$$

where $k > 0$ is the wavenumber (taken to be a constant) and the scattering function may be of compact support, i.e.

$$\gamma(x)\exists\forall x \in [-X,X].$$

The Forward Scattering Problem is defined as follows: Given $\gamma(x)\forall x$ find an exact solution for $u(x,k)$ The Inverse Scattering Problem is defined as follows: Given $u(x,k)\forall x$ find an exact solution for $\gamma(x)$.

APPENDIX 1.1.1 Theorem

Given that the (forward scattering Green's function) solution to the Helmholtz equation (as defined above) is

$$u(x,k) = u_i(x,k) + u_s(x,k) \tag{A1.1}$$

where u_i is a solution of

$$\left(\frac{\partial^2}{\partial x^2}+k^2\right)u_i(x,k) = 0,$$

$$u_s(x,k) = k^2 g(\mid x \mid,k) \otimes \gamma(x)u(x,k)$$

$$\equiv k^2 \int_{-\infty}^{\infty} g(\mid x-y \mid,k)\gamma(y)u(y,k)dy \tag{A1.2}$$

and

$$g(\mid x-y \mid,k) = \frac{i}{2k}\exp(ik \mid x-y \mid)$$

which is the (outgoing Green's function) solution of

$$\left(\frac{\partial^2}{\partial x^2}+k^2\right)g(\mid x-y \mid,k) = -\delta(x-y),$$

then

$$\gamma(x) = \frac{u^*(x,k)}{\mid u(x,k) \mid^2}\frac{\partial^2}{\partial x^2}\left[R(x)\otimes u_s(x,k) - \frac{1}{k^2}u_s(x,k)\right].$$

where (c_1 and c_2 being arbitrary constants)

$$R(x) = \begin{cases}(c_1-1)x+c_2, & x>0;\\ 0, & \text{otherwise.}\end{cases}$$

APPENDIX 1.1.2 Proof

From equations (A1.1) and (A1.2), we can write

$$(u - u_i) = k^2 g \otimes \gamma u.$$

Consider a piecewise continuous function q that is twice differentiable, such that

$$q \otimes (u - u_i) = k^2 q \otimes g \otimes \gamma u.$$

Differentiating twice, we have

$$\frac{\partial^2}{\partial x^2}[q \otimes (u - u_i)] = k^2 \frac{\partial^2}{\partial x^2}(q \otimes g \otimes \gamma u)$$

$$= k^2 \frac{\partial^2}{\partial x^2}(q \otimes g) \otimes \gamma u = -k^2 \delta \otimes \gamma u = -k^2 \gamma u$$

provided

$$\frac{\partial^2}{\partial x^2}(q \otimes g) = -\delta.$$

But

$$\frac{\partial^2}{\partial x^2}(q \otimes g) = q \otimes \frac{\partial^2}{\partial x^2} g$$

$$= q \otimes (-k^2 g - \delta) = -k^2 q \otimes g - q = -\delta$$

and hence

$$q = \delta - k^2 q \otimes g$$

so that

$$\frac{\partial^2}{\partial x^2}[q \otimes (u - u_i)] = \frac{\partial^2}{\partial x^2}[\delta \otimes (u - u_i) - k^2 q \otimes g \otimes (u - u_i)]$$

$$= \frac{\partial^2}{\partial x^2}[(u - u_i) - k^2 q \otimes g \otimes (u - u_i)] = -k^2 \gamma u.$$

Thus,

$$\gamma = \frac{1}{u} \frac{\partial^2}{\partial x^2} \left[q \otimes g \otimes (u - u_i) - \frac{1}{k^2}(u - u_i) \right]$$

The function q is determined by the solution of

$$\frac{\partial^2}{\partial x^2}(q \otimes g) = -\delta$$

$$\Longrightarrow \frac{\partial}{\partial x}(q \otimes g) = -H(x) + c_1 \qquad (A1.3)$$

where

$$H(x) = \begin{cases} 1, & x > 0; \\ 0, & \text{otherwise.} \end{cases}$$

and c_1 is a constant. But the solution of equation (A1.3) is

$$q \otimes g = R(x)$$

where (c_2 being a constant of integration)

$$R(x) = -\int_{-\infty}^{x} H(x)dx + c_1 x + c_2$$

$$= \begin{cases} (c_1 - 1)x + c_2, & x > 0; \\ 0, & \text{otherwise.} \end{cases}$$

so that

$$\gamma = \frac{1}{u}\frac{\partial^2}{\partial x^2}\left[R \otimes (u - u_i) - \frac{1}{k^2}(u - u_i)\right].$$

Finally, since $u = u_i + u_s$, we can write

$$\gamma = \frac{1}{u}\frac{\partial^2}{\partial x^2}\left[R \otimes u_s - \frac{1}{k^2}u_s\right]$$

$$= \frac{u^*}{|\ u\ |^2}\frac{\partial^2}{\partial x^2}\left[R \otimes u_s - \frac{1}{k^2}u_s\right]. \qquad (A1.4)$$

APPENDIX 1.1.3 Corollary

Since

$$\left(\frac{\partial^2}{\partial x^2} + k^2\right) u_i = 0.$$

it follows that (for $c_1 < 1$)

$$\begin{aligned} \gamma &= \frac{1}{u}\frac{\partial^2}{\partial x^2}\left[R \otimes (u - u_i) - \frac{1}{k^2}(u - u_i)\right] \\ &= \frac{1}{u}\left[-\delta \otimes (u - u_i) - \frac{1}{k^2}\left(\frac{\partial^2}{\partial x^2}u - \frac{\partial^2}{\partial x^2}u_i\right)\right] \\ &= \frac{1}{u}\left[-(u - u_i) - \frac{1}{k^2}\frac{\partial^2}{\partial x^2}u + \frac{1}{k^2}\frac{\partial^2}{\partial x^2}u_i\right] \\ &= \frac{1}{k^2 u}\left[-\left(\frac{\partial^2}{\partial x^2}u + k^2 u\right) + \frac{\partial^2}{\partial x^2}u_i + k^2 u_i\right] \\ &= -\frac{1}{k^2 u}\left(\frac{\partial^2}{\partial x^2} + k^2\right) u. \end{aligned}$$

which recovers the Helmholtz equation

$$\left(\frac{\partial^2}{\partial x^2} + k^2\right) u = -k^2 \gamma u.$$

APPENDIX 1.1.4 Remark I.1

The equivalent inverse solution for the Schrödinger equation

$$\left(\frac{\partial^2}{\partial x^2} + k^2\right) u(x,k) = \gamma(r) u(x,k)$$

is

$$\gamma = \frac{u^*}{\mid u \mid^2} \frac{\partial^2}{\partial x^2} \left[u_s - k^2 R \otimes u_s \right]. \qquad (A1.5)$$

APPENDIX 1.1.5 Remark I.2

The inverse solutions given by equations (A1.4) and (A1.5) rely on the condition:

$$\mid u(x,k) \mid = \mid u_i(x,k) + u_s(x,k) \mid > 0 \forall x, k.$$

Thus, for a fixed wavenumber $k > 0$, the incident u_i and scattered u_s wavefields must be 'out-of-phase' $\forall x$.

APPENDIX 1.1.6 Remark I.3

The theorem provides a result that is compatible with the trivial inverse solution

$$\gamma = -\frac{u^*}{\mid u \mid^2} \left(1 + \frac{1}{k^2} \frac{\partial^2}{\partial x^2} \right) u.$$

However, unlike this trivial solution, the theorem provides an expression for the scattering function γ which is, at least, consistent with the (exact) forward scattering (Green's function) solution $u = u_i + u_s$ and is determined by the scattered wavefield $u_s = k^2 g \otimes \gamma u$ that, like the incident wavefield u_i, is assumed to be a measurable quantity.

APPENDIX 1.1.7 Remark I.4

In order to use this inverse solution, the wavefield u must be known $\forall x$. For a scatterer of compact support, the field may only be measurable beyond this support and thus, the data on u may be incomplete.

APPENDIX 1.2 Exact Inverse Scattering Solution in Three-Dimensions

APPENDIX 1.2.1 Theorem

Consider the 3D inhomogeneous Helmholtz equation for a scalar (complex) wavefield $u(\mathbf{r}, k)$ given by

$$(\nabla^2 + k^2) u(\mathbf{r}, k) = -k^2 \gamma(\mathbf{r}) u(\mathbf{r}, k)$$

where k is the wavenumber and the scattering function is of compact support, i.e.

$$\gamma(\mathbf{r}) \exists \forall \mathbf{r} \in V.$$

Given that the forward (Green's function) solution to this equation is

$$u(\mathbf{r}, k) = u_i(\mathbf{r}, k) + u_s(\mathbf{r}, k) \qquad (A1.6)$$

where u_i is a solution of

$$(\nabla^2 + k^2)u_i(\mathbf{r}, k) = 0,$$

$$u_s(\mathbf{r}, k) = k^2 g(\mid \mathbf{r} \mid, k) \otimes_3 \gamma(\mathbf{r}) u(\mathbf{r}, k)$$

$$\equiv k^2 \int_V g(\mid \mathbf{r} - \mathbf{s} \mid, k)\gamma(\mathbf{s})u(\mathbf{s}, k)d^3\mathbf{s} \qquad (A1.7)$$

and

$$g(\mid \mathbf{r} - \mathbf{s} \mid, k) = \frac{\exp(ik \mid \mathbf{r} - \mathbf{s} \mid)}{4\pi \mid \mathbf{r} - \mathbf{s} \mid}$$

which is the solution of

$$(\nabla^2 + k^2)g(\mid \mathbf{r} - \mathbf{s} \mid, k) = -\delta^3(\mathbf{r} - \mathbf{s}),$$

then, with $r \equiv \mid \mathbf{r} \mid$,

$$\gamma(\mathbf{r}) = \frac{u^*(\mathbf{r}, k)}{\mid u(\mathbf{r}, k) \mid^2} \nabla^2 \left[\frac{1}{4\pi r} \otimes_3 u_s(\mathbf{r}, k) - \frac{1}{k^2} u_s(\mathbf{r}, k) \right].$$

APPENDIX 1.2.2 Proof

From equations (A1.6) and (A1.7), we can write

$$(u - u_i) = k^2 g \otimes_3 \gamma u.$$

Consider a function q such that

$$q \otimes_3 (u - u_i) = k^2 q \otimes_3 (g \otimes_3 \gamma u).$$

Taking the Laplacian of this equation, we have

$$\nabla^2[q \otimes_3 (u - u_i)] = k^2 \nabla^2 (q \otimes_3 g \otimes_3 \gamma u)$$

$$= k^2 \nabla^2 (q \otimes_3 g) \otimes_3 \gamma u = -k^2 \delta^3 \otimes_3 \gamma u = -k^2 \gamma u$$

provided

$$\nabla^2 (q \otimes_3 g) = -\delta^3.$$

But

$$\nabla^2 (q \otimes_3 g) = q \otimes_3 \nabla^2 g = q \otimes_3 (-k^2 g - \delta^3)$$

$$= -k^2 q \otimes_3 g - q = -\delta^3$$

and hence

$$q = \delta^3 - k^2 q \otimes_3 g$$

so that

$$\nabla^2[q \otimes_3 (u - u_i)] = \nabla^2[\delta^3 \otimes_3 (u - u_i) - k^2 q \otimes_3 g \otimes_3 (u - u_i)]$$

$$= \nabla^2[(u - u_i) - k^2 q \otimes_3 g \otimes_3 (u - u_i)] = -k^2 \gamma u.$$

Thus,

$$\gamma = \frac{1}{u}\nabla^2 \left[q \otimes_3 g \otimes_3 (u - u_i) - \frac{1}{k^2}(u - u_i) \right]$$

where q is determined by the solution of

$$\nabla^2 (q \otimes_3 g) = -\delta^3. \qquad (A1.8)$$

But the solution of equation (A1.8) is

$$q \otimes_3 g = \frac{1}{4\pi r}$$

so that

$$\gamma = \frac{1}{u}\nabla^2 \left[\frac{1}{4\pi r} \otimes_3 (u - u_i) - \frac{1}{k^2}(u - u_i) \right].$$

Finally, since $u = u_i + u_s$, we can write

$$\gamma = \frac{1}{u}\nabla^2 \left[\frac{1}{4\pi r} \otimes_3 u_s - \frac{1}{k^2} u_s \right]$$

$$= \frac{u^*}{|\, u \,|^2}\nabla^2 \left[\frac{1}{4\pi r} \otimes_3 u_s - \frac{1}{k^2} u_s \right]. \qquad (A1.9)$$

APPENDIX 1.2.3 Corollary

Since

$$(\nabla^2 + k^2) u_i = 0.$$

it follows that

$$\begin{aligned}
\gamma &= \frac{1}{u}\nabla^2 \left[\frac{1}{4\pi r} \otimes_3 (u - u_i) - \frac{1}{k^2}(u - u_i) \right] \\
&= \frac{1}{u} \left[-\delta^3 \otimes_3 (u - u_i) - \frac{1}{k^2}(\nabla^2 u - \nabla^2 u_i) \right] \\
&= \frac{1}{u} \left[-(u - u_i) - \frac{1}{k^2}\nabla^2 u + \frac{1}{k^2}\nabla^2 u_i \right] \\
&= \frac{1}{k^2 u} \left[-(\nabla^2 u + k^2 u) + \nabla^2 u_i + k^2 u_i \right] \\
&= -\frac{1}{k^2 u}(\nabla^2 + k^2) u.
\end{aligned}$$

which recovers the Helmholtz equation

$$(\nabla^2 + k^2) u = -k^2 \gamma u.$$

APPENDIX 1.2.4 Remark II.1

Remarks I.1-I.4 apply to this three dimensional derivation.

APPENDIX 1.2.5 Remark II.2

In the 2D case, equation (A1.9) becomes

$$\nabla^2(q \otimes_2 g) = -\delta^2$$

and has the solution

$$q \otimes_2 g = \frac{1}{2\pi} \ln r$$

where $\otimes_2$ denotes the two-dimensional convolution integral. The equivalent 2D inverse solution is then given by

$$\gamma(\mathbf{r}) = \frac{u^*(\mathbf{r},k)}{\mid u(\mathbf{r},k) \mid^2} \nabla^2 \left[\frac{1}{2\pi} \ln r \otimes_2 u_s(\mathbf{r},k) - \frac{1}{k^2} u_s(\mathbf{r},k) \right].$$

APPENDIX 1.2.6 Remark II.3

Equation A2.9 relies on the boundary condition $u(\mathbf{r},,k) = u_i(\mathbf{r},k) \ \ \forall \mathbf{r} \in S$ where S defines the surface of $\gamma(\mathbf{r})$ which is taken to be of compact support. The Green's function solution to the three dimensional inhomogeneous Helmholtz equation is

$$u(\mathbf{r},k) = k^2 \int_V g\gamma u d^3\mathbf{s} + \oint_S (g\nabla u - u\nabla g) \cdot \hat{\mathbf{n}} d^2\mathbf{s}.$$

To compute the surface integral, a condition for the behaviour of u on the surface S of γ must be chosen. If we consider the case where the incident wavefield u_i is a simple plane wave of unit amplitude

$$\exp(i\mathbf{k} \cdot \mathbf{r})$$

satisfying the homogeneous wave equation

$$(\nabla^2 + k^2) u_i(\mathbf{r},k) = 0,$$

then

$$u(\mathbf{r},k) = k^2 \int_V g\gamma u d^3\mathbf{s} + \oint_S (g\nabla u_i - u_i \nabla g) \cdot \hat{\mathbf{n}} d^2\mathbf{s}.$$

Using Green's theorem to convert the surface integral back into a volume integral, we have

$$\oint_S (g\nabla u_i - u_i \nabla g) \cdot \hat{\mathbf{n}} d^2\mathbf{s} = \int_V (g\nabla^2 u_i - u_i \nabla^2 g) d^3\mathbf{s}.$$

Noting that

$$\nabla^2 u_i = -k^2 u_i$$

and that

$$\nabla^2 g = -\delta^3 - k^2 g$$

we obtain

$$\int_V (g\nabla^2 u_i - u_i \nabla^2 g) d^3\mathbf{s} = \int \delta^3 u_i d^3\mathbf{s} = u_i.$$

Hence, by choosing the field u to be equal to the incident wavefield u_i on the surface of γ, we obtain a solution of the form

$$u = u_i + u_s$$

where

$$u_s(\mathbf{r}, k) = k^2 g(\mid \mathbf{r} \mid, k) \otimes_3 \gamma(\mathbf{r}) u(\mathbf{r}, k).$$

APPENDIX 2 RELATIONSHIP BETWEEN THE HURST EXPONENT AND THE TOPOLOGICAL, FRACTAL AND FOURIER DIMENSIONS

Suppose we cut up some simple one-, two- and three-dimensional Euclidean objects (a line, a square surface and a cube, for example), make exact copies of them and then keep on repeating the copying process. Let N be the number of copies that we make at each stage and let r be the length of each of the copies, i.e. the scaling ratio. Then we have

$$Nr^{D_T} = 1, \quad D_T = 1, 2, 3, ...$$

where D_T is the topological dimension. The similarity or fractal dimension is that value of D_F which is usually (but not always) a non-integer dimension 'greater' that its topological dimension (i.e. 0,1,2,3,... where 0 is the dimension of a point on a line) and is given by

$$D_F = -\frac{\log(N)}{\log(r)}.$$

The fractal dimension is that value that is strictly greater than the topological dimension as given in Table II. In each case, as the value of the fractal dimension

TABLE 6 Fractal types and corresponding fractal dimensions

Fractal type	Fractal Dimension
Fractal Dust	$0 < D_F < 1$
Fractal Curve	$1 < D_F < 2$
Fractal Surface	$2 < D_F < 3$
Fractal Volume	$3 < D_F < 4$
Fractal Time	$4 < D_F < 5$
Hyper-fractals	$5 < D_F < 6$
⋮	⋮

increases, the fractal becomes increasingly 'space-filling' in terms of the topological dimension which the fractal dimension is approaching. In each case, the fractal exhibits structures that are self-similar. A self-similar deterministic fractal is one where a change in the scale of a function $f(x)$ (which may be a multi-dimensional function) by a scaling factor λ produces a smaller version, reduced in size by λ, i.e.

$$f(\lambda x) = \lambda f(x).$$

A self-affine deterministic fractal is one where a change in the scale of a function $f(x)$ by a factor λ produces a smaller version reduced in size by a factor λ^q, $q > 0$, i.e.

$$f(\lambda x) = \lambda^q f(x).$$

For stochastic fields, the expression

$$\Pr[f(\lambda x)] = \lambda^q \Pr[f(x)]$$

describes a statistically self-affine field - a random scaling fractal. As we zoom into the fractal, the shape changes, but the distribution of lengths remains the same.

There is no unique method for computing the fractal dimension. The methods available are broadly categorized into two families: (i) Size-measure relationships, based on recursive length or area measurements of a curve or surface using different measuring scales; (ii) application of relationships based on approximating or fitting a curve or surface to a known fractal function or statistical property, such as the variance.

Consider a simple Euclidean straight line ℓ of length $L(\ell)$ over which we 'walk' a shorter 'ruler' of length δ. The number of steps taken to cover the line $N[L(\ell),\delta]$ is then L/δ which is not always an integer for arbitrary L and δ. Since

$$N[L(\ell),\delta] = \frac{L(\ell)}{\delta} = L(\ell)\delta^{-1},$$

$$\Rightarrow 1 = \frac{\ln L(\ell) - \ln N[L(\ell),\delta]}{\ln \delta} = -\left(\frac{\ln N[L(\ell),\delta] - \ln L(\ell)}{\ln \delta}\right)$$

which expresses the topological dimension $D_T = 1$ of the line. In this case, $L(\ell)$ is the Lebesgue measure of the line and if we normalize by setting $L(\ell) = 1$, the latter equation can then be written as

$$1 = -\lim_{\delta \to 0}\left[\frac{\ln N(\delta)}{\ln \delta}\right]$$

since there is less error in counting $N(\delta)$ as δ becomes smaller. We also then have $N(\delta) = \delta^{-1}$. For extension to a fractal curve f, the essential point is that the fractal dimension should satisfy an equation of the form

$$N[F(f),\delta] = F(f)\delta^{-D_F}$$

where $N[F(f),\delta]$ is 'read' as the number of rulers of size δ needed to cover a fractal set f whose measure is $F(f)$ which can be any valid suitable measure of the curve. Again we may normalize, which amounts to defining a new measure F' as some constant multiplied by the old measure to get

$$D_F = -\lim_{\delta \to 0}\left[\frac{\ln N(\delta)}{\ln \delta}\right]$$

where $N(\delta)$ is taken to be $N[F'(f),\delta]$ for notational convenience. Thus a piece-wise continuous field has precise fractal properties over all scales. However, for the discrete (sampled) field

$$D = -\left\langle \frac{\ln N(\delta)}{\ln \delta} \right\rangle$$

where we choose values δ_1 and δ_2 (i.e. the upper and lower bounds) satisfying $\delta_1 < \delta < \delta_2$ over which we apply an averaging processes denoted by $\langle\ \rangle$. The most common approach is to utilise a bi-logarithmic plot of $\ln N(\delta)$ against $\ln \delta$, choose values δ_1 and δ_2 over which the plot is uniform and apply an appropriate

data fitting algorithm (e.g. a least squares estimation method or, as used in this paper, Orthogonal Linear Regression) within these limits.

The relationship between the Fourier dimension q and the fractal dimension D_F can be determined by considering this method for analysing a statistically self-affine field. For a fractional Brownian process (with unit step length)

$$A(t) = t^H, \;\; H \in (0, 1]$$

where H is the Hurst dimension. Consider a fractal curve covering a time period $\Delta t = 1$ which is divided up into $N = 1/\Delta t$ equal intervals. The amplitude increments ΔA are then given by

$$\Delta A = \Delta t^H = \frac{1}{N^H} = N^{-H}.$$

The number of lengths $\delta = N^{-1}$ required to cover each interval is

$$\Delta A \Delta t = \frac{N^{-H}}{N^{-1}} = N^{1-H}$$

so that

$$N(\delta) = N N^{1-H} = N^{2-H}.$$

Now, since

$$N(\delta) = \frac{1}{\delta^{D_F}}, \;\; \delta \to 0,$$

then, by inspection,

$$D_F = 2 - H.$$

Thus, a Brownian process, where $H = 1/2$, has a fractal dimension of 1.5. For higher topological dimensions D_T

$$D_F = D_T + 1 - H.$$

This algebraic equation provides the relationship between the fractal dimension D_F, the topological dimension D_T and the Hurst dimension H. We can now determine the relationship between the Fourier dimension q and the fractal dimension D_F.

Consider a fractal signal $f(x)$ over an infinite support with a finite sample $f_X(x)$, given by

$$f_X(x) = \begin{cases} f(x), & 0 < x < X; \\ 0, & \text{otherwise.} \end{cases}$$

A finite sample is essential as otherwise the power spectrum diverges. Moreover, if $f(x)$ is a random function then for any experiment or computer simulation we must necessarily take a finite sample. Let $F_X(k)$ be the Fourier transform of $f_X(x)$, $P_X(k)$ be the power spectrum and $P(k)$ be the power spectrum of $f(x)$. Then

$$f_X(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} F_X(k) \exp(ikx) dk,$$

$$P_X(k) = \frac{1}{X} |F_X(k)|^2$$

and

$$P(k) = \lim_{X \to \infty} P_X(k).$$

The power spectrum gives an expression for the power of a signal for particular harmonics. $P(k)dk$ gives the power in the range k to $k + dk$. Consider a function $g(x)$, obtained from $f(x)$ by scaling the x-coordinate by some $a > 0$, the f-coordinate by $1/a^H$ and then taking a finite sample as before, i.e.

$$g_X(x) = \begin{cases} g(x) = \frac{1}{a^H} f(ax), & 0 < x < X; \\ 0, & \text{otherwise.} \end{cases}$$

Let $G_X(k)$ and $P'_X(k)$ be the Fourier transform and power spectrum of $g_X(x)$, respectively. We then obtain an expression for G_X in terms of F_X,

$$G_X(k) = \int_0^X g_X(x) \exp(-ikx) dx =$$

$$\frac{1}{a^{H+1}} \int_0^X f(s) \exp\left(-\frac{iks}{a}\right) ds$$

where $s = ax$. Hence

$$G_X(k) = \frac{1}{a^{H+1}} F_X\left(\frac{k}{a}\right)$$

and the power spectrum of $g_X(x)$ is

$$P'_X(k) = \frac{1}{a^{2H+1}} \frac{1}{aX} \left| F_X\left(\frac{k}{a}\right) \right|^2$$

and, as $X \to \infty$,

$$P'(k) = \frac{1}{a^{2H+1}} P\left(\frac{k}{a}\right).$$

Since $g(x)$ is a scaled version of $f(x)$, their power spectra are equal, and so

$$P(k) = P'(k) = \frac{1}{a^{2H+1}} P\left(\frac{k}{a}\right).$$

If we now set $k = 1$ and then replace $1/a$ by k we get

$$P(k) \propto \frac{1}{k^{2H+1}} = \frac{1}{k^\beta}.$$

Now since $\beta = 2H + 1$ and $D_F = 2 - H$, we have

$$D_F = 2 - \frac{\beta - 1}{2} = \frac{5 - \beta}{2}.$$

The fractal dimension of a fractal signal can be calculated directly from β using the above relationship. This method also generalizes to higher topological dimensions giving

$$\beta = 2H + D_T.$$

Thus, since

$$D_F = D_T + 1 - H,$$

then $\beta = 5 - 2D_F$ for a fractal signal and $\beta = 8 - 2D_F$ for a fractal surface so that, in general,

$$\beta = 2(D_T + 1 - D_F) + D_T = 3D_T + 2 - 2D_F$$

and

$$D_F = D_T + 1 - H = D_T + 1 - \frac{\beta - D_T}{2} = \frac{3D_T + 2 - \beta}{2},$$

the Fourier dimension being given by $q = \beta/2$.

APPENDIX 3 OVERVIEW OF FRACTIONAL CALCULUS

Fractional calculus (e.g. [84], [85], [86], [87], [88]) is the study of the calculus associated with fractional differentials and a fractional integrals which, in the main, are based on generalizations of results obtained using integer calculus. For example, the classical fractional integral operators are the Riemann-Liouville transform [84]

$$\hat{I}^q f(t) = \frac{1}{\Gamma(q)} \int_{-\infty}^{t} \frac{f(\tau)}{(t-\tau)^{1-q}} d\tau, \quad q > 0$$

and the Weyl transform

$$\hat{I}^q f(t) = \frac{1}{\Gamma(q)} \int_{t}^{\infty} \frac{f(\tau)}{(t-\tau)^{1-q}} d\tau, \quad q > 0$$

where

$$\Gamma(q) = \int_{0}^{\infty} t^{q-1} \exp(-t) dt.$$

For integer values of q (i.e. when $q = n$ where n is a non-negative integer), the Riemann-Liouville transform reduces to the standard Riemann integral. This transform is just a (causal) convolution of the function $f(t)$ with $t^{q-1}/\Gamma(q)$. For fractional differentiation, we can perform a fractional integration of appropriate order and then differentiate to an appropriate integer order. The reason for this is that direct fractional differentiation can lead to divergent integrals. Thus, the fractional differential operator $\hat{D}^q$ for $q > 0$ is given by

$$\hat{D}^q f(t) \equiv \frac{d^q}{dt^q} f(t) = \frac{d^n}{dt^n} [\hat{I}^{n-q} f(t)].$$

where

$$\hat{I}^{q-n} f(t) = \frac{1}{\Gamma(n-q)} \int_{-\infty}^{t} \frac{f(\tau)}{(t-\tau)^{1+q-n}} d\tau, \quad n - q > 0$$

in which the value of $\hat{I}^{q-n} f(t)$ at a point t depends on the behaviour of $f(t)$ from $-\infty$ to t via a convolution with the kernel $t^{n-q}/\Gamma(q)$. The convolution process is dependent on the history of the function $f(t)$ for a given kernel and thus, in this context, we can consider a fractional derivative defined via the result above to have memory.

APPENDIX 3.1 The Laplace Transform and the Half Integrator

It informative at this point to consider the application of the Laplace transform to identify an ideal integrator and then a half integrator. The Laplace transform is

given by

$$\hat{L}[f(t)] \equiv F(p) = \int_0^\infty f(t)\exp(-pt)dt$$

and from this result we can derive the transform of a derivative given by

$$\hat{L}[f'(t)] = pF(p) - f(0)$$

and the transform of an integral given by

$$\hat{L}\left[\int_0^t f(\tau)d\tau\right] = \frac{1}{p}F(p).$$

Now, suppose we have a standard time invariant linear system whose input is $f(t)$ and whose output is given by

$$s(t) = f(t) \otimes g(t)$$

where the convolution is causal, i.e.

$$s(t) = \int_0^t f(\tau)g(t-\tau)d\tau.$$

Suppose we let

$$g(t) = H(t) = \begin{cases} 1, & t > 0; \\ 0, & t < 0. \end{cases}$$

Then, $G(p) = 1/p$ and the system becomes an ideal integrator:

$$s(t) = f(t) \otimes H(t) = \int_0^t f(t-\tau)d\tau = \int_0^t f(\tau)d\tau.$$

Now, consider the case when we have a time invariant linear system with an impulse response function by given by

$$g(t) = \frac{H(t)}{\sqrt{t}} = \begin{cases} \mid t \mid^{-1/2}, & t > 0; \\ 0, & t < 0. \end{cases}$$

The output of this system is $f \otimes g$ and the output of such a system with input $f \otimes g$ is $f \otimes g \otimes g$. Now

$$g(t) \otimes g(t) = \int_0^t \frac{d\tau}{\sqrt{\tau}\sqrt{t-\tau}} = \int_0^{\sqrt{t}} \frac{2xdx}{x\sqrt{t-x^2}}$$

$$= 2\left[\sin^{-1}\left(\frac{x}{\sqrt{t}}\right)\right]_0^{\sqrt{t}} = \pi.$$

Hence,

$$\frac{H(t)}{\sqrt{\pi t}} \otimes \frac{H(t)}{\sqrt{\pi t}} = H(t)$$

and the system defined by the impulse response function $H(t)/\sqrt{\pi t}$ represents a 'half-integrator' with a Laplace transform given by

$$\hat{L}\left[\frac{H(t)}{\sqrt{\pi t}}\right] = \frac{1}{\sqrt{p}}.$$

This result provides an approach to working with fractional integrators and/or differentiators using the Laplace transform. Fractional differential and integral operators can be defined and used in a similar manner to those associated with conventional or integer order calculus and we now provide an overview of such operators.

APPENDIX 3.2 Operators of Integer Order

The following operators are all well-defined, at least with respect to all test functions $u(t)$ say which are (i) infinitely differentiable and (ii) of compact support (i.e. vanish outside some finite interval).

Integral Operator:

$$\hat{I}u(t) \equiv \hat{I}^1 u(t) = \int_{-\infty}^{t} u(\tau)d\tau.$$

Differential Operator:

$$\hat{D}u(t) \equiv \hat{D}^1 u(t) = u'(t).$$

Identify Operator:

$$\hat{I}^0 u(t) = u(t) = \hat{D}^0 u(t).$$

Now,

$$\hat{I}[\hat{D}u](t) = \int_{-\infty}^{t} u'(\tau)d\tau = u(t)$$

and

$$\hat{D}[\hat{I}u](t) = \frac{d}{dt}\int_{-\infty}^{t} u(\tau)d\tau = u(t)$$

so that

$$\hat{I}^1\hat{D}^1 = \hat{D}^1\hat{I}^1 = \hat{I}^0.$$

For n (integer) order:

$$\hat{I}^n u(t) = \int_{-\infty}^{t} d\tau_{n-1} ... \int_{-\infty}^{\tau_2} d\tau_1 \int_{-\infty}^{\tau_1} u(\tau) d\tau,$$

$$\hat{D}^n u(t) = u^{(n)}(t)$$

and

$$\hat{I}^n[\hat{D}^n u](t) = u(t) = \hat{D}^n[\hat{I}^n u](t).$$

APPENDIX 3.3 Convolution Representation

Consider the function

$$t_+^{q-1}(t) \equiv \mid t \mid^{q-1} H(t) = \begin{cases} \mid t \mid^{q-1}, & t > 0; \\ 0, & t < 0. \end{cases}$$

which, for any $q > 0$ defines a function that is locally integrable. We can then define an integral of order n in terms of a convolution as

$$\hat{I}^n u(t) = \left(u \otimes \frac{1}{(n-1)!} t_+^{n-1} \right)(t)$$

$$= \frac{1}{(n-1)!} \int_{-\infty}^{t} (t-\tau)^{n-1} u(\tau) d\tau$$

$$= \frac{1}{(n-1)!} \int_{-\infty}^{t} \tau^{n-1} u(t-\tau) d\tau$$

In particular,

$$\hat{I}^1 u(t) = (u \otimes H)(t) = \int_{-\infty}^{t} u(\tau) d\tau.$$

These are classical (absolutely convergent) integrals and the identity operator admits a formal convolution representation, using the delta function, i.e.

$$\hat{I}^0 u(t) = \int_{-\infty}^{\infty} \delta(\tau) u(t-\tau) d\tau$$

where

$$\delta(t) = \hat{D} H(t).$$

Similarly,

$$\hat{D}^n u(t) \equiv \hat{I}^{-n} u(t) = \int_{-\infty}^{\infty} \delta^{(n)}(\tau) u(t-\tau) d\tau = u^{(n)}(t).$$

On the basis of the material discussed above, we can now formally extend the integral operator to fractional order and consider the operator

$$\hat{I}^q u(t) = \frac{1}{\Gamma(q)} \int_{-\infty}^{\infty} u(\tau) t_+^{q-1} (t-\tau) d\tau$$

$$= \frac{1}{\Gamma(q)} \int_{-\infty}^{t} u(\tau) t_+^{q-1} (t-\tau) d\tau$$

where

$$\Gamma(q) = \int_0^{\infty} t^{q-1} \exp(-t) dt, \quad q > 0$$

with the fundamental property that

$$\Gamma(q+1) = q\Gamma(q).$$

Here, I^q is an operator representing a time invariant linear system with impulse response function $t_+^{q-1}(t)$ and transfer function $1/p^q$. For the cascade connection of I^{q_1} and I^{q_2} we have

$$\hat{I}^{q_1}[\hat{I}^{q_2} u(t)] = \hat{I}^{q_1+q_2} u(t).$$

This classical convolution integral representation holds for all real $q > 0$ (and formally for $q = 0$, with the delta function playing the role of an impulse function and with a transfer function equal to the constant 1).

APPENDIX 3.4 Fractional Differentiation

For $0 < q < 1$, if we define the (Riemann-Liouville) derivative of order q as

$$\hat{D}^q u(t) \equiv \frac{d}{dt}[\hat{I}^{1-q} u](t) = \frac{1}{\Gamma(1-q)} \frac{d}{dt} \int_{-\infty}^{t} (t-\tau)^{-q} u(\tau) d\tau,$$

then,

$$\hat{D}^q u(t) = \frac{1}{\Gamma(1-q)} \int_{-\infty}^{t} (t-\tau)^{-q} u'(\tau) d\tau \equiv \hat{I}^{1-q} u'(t).$$

Hence,

$$\hat{I}^q[\hat{D}^q u] = \hat{I}^q[\hat{I}^{1-q} u'] = \hat{I}^1 u' = u$$

and $\hat{D}^q$ is the formal inverse of the operator $\hat{I}^q$. Given any $q > 0$, we can always write $\lambda = n - 1 + q$ and then define

$$\hat{D}^\lambda u(t) = \frac{1}{\Gamma(1-q)} \frac{d^n}{dt^n} \int_{-\infty}^{t} u(\tau)(t-\tau)^{-q} d\tau.$$

D^q is an operator representing a time invariant linear system consisting of a cascade combination of an ideal differentiator and a fractional integrator of order $1-q$. For D^λ we replace the single ideal differentiator by n such that

$$\hat{D}^0 u(t) = \frac{1}{\Gamma(1)} \frac{d}{dt} \int_{-\infty}^{t} u(\tau) d\tau = u(t) \equiv \int_{-\infty}^{\infty} u(\tau)\delta(t-\tau)d\tau$$

and

$$\hat{D}^n u(t) = \frac{1}{\Gamma(1)} \frac{d^{n+1}}{dt^{n+1}} \int_{-\infty}^{t} u(\tau) d\tau$$

$$= u^{(n)}(t) \equiv \int_{-\infty}^{\infty} u(\tau)\delta^{(n)}(t-\tau)d\tau.$$

In addition to the conventional and classical definitions of fractional derivatives and integrals, more general definitions are available including the Erdélyi-Kober fractional integral [89]

$$\frac{t^{-p-q+1}}{\Gamma(q)} \int_0^t \frac{\tau^{p-1}}{(t-\tau)^{1-q}} f(\tau)d\tau, \quad q > 0, \quad p > 0$$

which is a generalisation of the Riemann-Liouville fractional integral and the integral

$$\frac{t^p}{\Gamma(q)} \int_t^\infty \frac{\tau^{-q-p}}{(\tau-t)^{1-q}} f(\tau)d\tau, \quad q > 0, \quad p > 0$$

which is a generalization of the Weyl integral. Further definitions exist based on the application of hypergeometric functions and operators involving other special functions such as the Maijer G-function and the Fox H-function [88]. Moreover, all such operators leading to a fractional integral of the Riemann-Liouville type and the Weyl type to have the general forms (through induction)

$$\hat{I}^q f(t) = t^{q-1} \int_{-\infty}^{t} \Phi\left(\frac{\tau}{t}\right) \tau^{-q} f(\tau)d\tau$$

and

$$\hat{I}^q f(t) = t^{-q} \int_{t}^{\infty} \Phi\left(\frac{t}{\tau}\right) \tau^{q-1} f(\tau)d\tau$$

respectively, where the kernel Φ is an arbitrary continuous function so that the integrals above make sense in sufficiently large functional spaces. Although there are a number of approaches that can be used to define a fractional differential/integral,

there is one particular definition, which has wide ranging applications in signal and image processing and is based on the Fourier transform, i.e.

$$\frac{d^q}{dt^q} f(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} (i\omega)^q F(\omega) \exp(i\omega t) d\omega$$

where $F(\omega)$ is the Fourier transform of $f(t)$.

APPENDIX 4 SCALING LAW FOR A RANDOM SELF-AFFINE FUNCTION

Self-affine functions are characterised by an amplitude spectral density function of the type k^{-q} where $k = \mid \mathbf{k} \mid$ is the spatial frequency. This appendix provides detail on calculating the n-dimensional (inverse) Fourier transform of such a spectrum which is compounded in the following theorem:

Theorem If $q \neq 2m$ or $-n-2m$ where $m = 0, 1, 2, ...$, then

$$\mathcal{F}_n[r^q] = \int_{-\infty}^{\infty} r^q \exp(-i\mathbf{k}\cdot\mathbf{r}) d^n\mathbf{r} = \frac{(\frac{1}{2}q + \frac{1}{2}n - 1)!}{(-\frac{1}{2}q - 1)!} 2^{q+n} \pi^{n/2} k^{-q-n}$$

where $\mathbf{k}$ and $\mathbf{r}$ are the n-dimensional vectors $(k_1, k_2, ..., k_n)$ and $(r_1, r_2, ..., r_n)$ respectively, $r \equiv \mid \mathbf{r} \mid = \sqrt{r_1^2 + r_2^2 + ... + r_n^2}$ and $k \equiv \mid \mathbf{k} \mid = \sqrt{k_1^2 + k_2^2 + ... + k_n^2}$. Note that

$$\mathcal{F}_n[f(\mathbf{r})] = \int_{-\infty}^{\infty} f(\mathbf{r}) \exp(-i\mathbf{k}\cdot\mathbf{r}) d^n\mathbf{r}$$

is taken to mean

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} f(r_1, r_2, ..., r_n) \exp[-i(k_1 r_1 + k_2 r_2 + ... + k_n r_n)] dr_1 dr_2 ..., dr_n.$$

Proof The proof of this result is based two results:

(i) If f is a function of r only, then

$$F(\mathbf{k}) = \left(1 - \frac{\partial^2}{\partial k_1^2} - \frac{\partial^2}{\partial k_2^2} - ... - \frac{\partial^2}{\partial k_n^2}\right)^N (2\pi)^{n/2} \int_0^{\infty} \frac{f(r) r^{n-1}}{(1+r^2)^N} \frac{J_{\frac{n-2}{2}}(kr)}{(kr)^{(n/2)-1}} dr$$

where N is a positive integer and $J_{(n-2)/2}$ is the Bessel function (of order $(n-2)/2$).

(ii) For Bessel Functions,

$$\frac{(2\pi)^{n/2}}{k^{(n/2)-1}} \int_0^{\infty} \frac{r^{q+(n/2)}}{(1+r^2)^N} J_{\frac{n-2}{2}}(kr) dr$$

$$= \frac{\pi^{n/2}(\frac{1}{2}q + \frac{1}{2}n - 1)!(N - \frac{1}{2}q - \frac{1}{2}n - 1)!}{(N-1)!(\frac{1}{2}n - 1)!} {}_1F_2(\tfrac{1}{2}q + \tfrac{1}{2}n; \tfrac{1}{2}q + \tfrac{1}{2}n - N + 1, \tfrac{1}{2}n; \tfrac{1}{4}k^2)$$

$$+ \frac{\pi^{n/2} k^{2N-q-n} (\frac{1}{2}q + \frac{1}{2}n - N - 1)!}{(N - \frac{1}{2}q - 1)! 2^{2N-q-n}} {}_1F_2(N; N - \tfrac{1}{2}q, N + 1 - \tfrac{1}{2}q - \tfrac{1}{2}n; \tfrac{1}{4}k^2) \quad (A4.1)$$

where

$${}_1F_2(a; b, c; x) = 1 + \frac{a}{1!bc}x + \frac{a(a+1)}{2!b(b+1)c(c+1)}x^2 + ...$$

The first of these results can be obtained by choosing a polar axis to lie along the direction of $\mathbf{k}$ so that $\mathbf{k}\cdot\mathbf{r} = kr\cos\theta_1$ and

$$F(\mathbf{k}) = \int_{-\infty}^{\infty} f(r)\exp(-i\mathbf{k}\cdot\mathbf{r})d\mathbf{r} = \int_0^{\infty} f(r)r^{n-1}\int_0^{\pi}\exp(-ikr\cos\theta_1)\sin^{n-2}\theta_1 d\theta_1$$

$$\times\int_0^{\pi}\ldots\int_0^{2\pi}\sin^{n-3}\theta_2\ldots\sin\theta_{n-2}d\theta_2\ldots d\theta_{n-1}dr$$

$$= \int_0^{\infty} f(r)r^{n-1}\frac{2\pi^{(n-1)/2}}{(\frac{1}{2}n-\frac{3}{2})!}\int_0^{\pi}\exp(-ikr\cos\theta_1)\sin^{n-2}\theta_1 d\theta_1 dr$$

using

$$\int_0^{\pi}\sin^{\nu}d\theta = \frac{(\frac{1}{2}\nu-\frac{1}{2})!\pi^{1/2}}{(\frac{1}{2}\nu)!}.$$

Now,

$$-\left(\frac{\partial^2}{\partial k_1^2}+\frac{\partial^2}{\partial k_2^2}+\ldots+\frac{\partial^2}{\partial k_n^2}\right) = \int_{-\infty}^{\infty} f(r)(r_1^2+r_2^2+\ldots+r_n^2)\exp(-i\mathbf{k}\cdot\mathbf{r})d^n\mathbf{r}$$

and therefore

$$\left(1-\frac{\partial^2}{\partial k_1^2}-\frac{\partial^2}{\partial k_2^2}-\ldots-\frac{\partial^2}{\partial k_n^2}\right)^N = \int_{-\infty}^{\infty} f(r)(1+r^2)^N\exp(-i\mathbf{k}\cdot\mathbf{r})d^n\mathbf{r}.$$

Hence, we can write

$$F(\mathbf{k}) = \left(1-\frac{\partial^2}{\partial k_1^2}+\frac{\partial^2}{\partial k_2^2}-\ldots-\frac{\partial^2}{\partial k_n^2}\right)^N(2\pi)^{n/2}\int_0^{\infty}\frac{f(r)r^{n-1}}{(1+r^2)^N}\frac{J_{\frac{n-2}{2}}(kr)}{(kr)^{(n/2)-1}}dr. \tag{A4.2}$$

The ratio of two successive terms u_{n+1}/u_n in the infinite series for ${}_1F_2$ is $(a+n)x/[(n+1)(b+n)(c+n)]$ which tends to zero as $n\to\infty$ for any finite x. Thus, the series for ${}_1F_2$ converges absolutely and uniformly with respect to x and the same is true of its derivatives (provided that neither b or c is a negative integer or zero when the series diverges). Therefore,

$$\left(\frac{\partial^2}{\partial k_1^2}+\frac{\partial^2}{\partial k_2^2}+\ldots+\frac{\partial^2}{\partial k_n^2}\right){}_1F_2(a;b,\tfrac{1}{2}n;\tfrac{1}{4}k^2)$$

$$= \frac{(b-1)!(\frac{1}{2}-1)!}{(a-1)!}\left(\frac{\partial^2}{\partial k_1^2}+\frac{\partial^2}{\partial k_2^2}+\ldots+\frac{\partial^2}{\partial k_n^2}\right)\sum_{s=0}^{\infty}\frac{(a+s-1)!(\frac{1}{2}k)^{2s}}{(b+s-1)!(\frac{1}{2}n+s-1)!s!}$$

$$= \frac{(b-1)!(\frac{1}{2}n-1)!}{(a-1)!}\sum_{s=0}^{\infty}\frac{(a+s-1)!(\frac{1}{2}k)^{2s-2}}{(b+s-1)!(\frac{1}{2}n+s-2)!(s-1)!}.$$

The term for $s = 0$ disappears so that, by replacing s by $s+1$ we obtain

$$\left(\frac{\partial^2}{\partial k_1^2}+\frac{\partial^2}{\partial k_2^2}+...+\frac{\partial^2}{\partial k_n^2}-1\right){}_1F_2(a;b,\tfrac{1}{2}n;\tfrac{1}{4}k^2)$$

$$=\frac{(b-1)!(\frac{1}{2}n-1)!}{(a-1)!}\sum_{s=0}^{\infty}\frac{(a+s-1)!(\frac{1}{2}k)^{2s}}{(b+s)!(\frac{1}{2}n+s-1)!s!}(a+s-b-s)$$

$$=\frac{a-b}{b}{}_1F_2(a;b+1,\tfrac{1}{2}n;\tfrac{1}{4}k^2).$$

Hence,

$$\left(\frac{\partial^2}{\partial k_1^2}+\frac{\partial^2}{\partial k_2^2}+...+\frac{\partial^2}{\partial k_n^2}-1\right)^N{}_1F_2(a;b,\tfrac{1}{2}n;\tfrac{1}{4}k^2)$$

$$=\frac{(a-b)(a-b-1)...(a-b-N+1)}{b(b+1)...(b+N-1)}{}_1F_2(a;b+N,\tfrac{1}{2}n;\tfrac{1}{4}k^2). \qquad (A4.3)$$

In the first term of equation (A4.1) $a=\frac{1}{2}(q+n), b=\frac{1}{2}(q+n)-N+1$ so that $a-b=N+1$ with the result that the right hand side of the equation vanishes. For the second term of equation (A4.1), consider, with $b>0$

$$\left(\frac{\partial^2}{\partial k_1^2}+\frac{\partial^2}{\partial k_2^2}+...+\frac{\partial^2}{\partial k_n^2}\right)k^{2b}{}_1F_2(a;b+\tfrac{1}{2}n,b+1;\tfrac{1}{4}k^2)$$

$$=\frac{(b+\frac{1}{2}n-1)!b!}{(a-1)!}\sum_{s=0}^{\infty}\frac{(a+s-1)!k^{2b+2s-2}}{4^{s-1}(b+\frac{1}{2}n-2+s)!(b+s-1)!s!}$$

as above. Hence,

$$\left(\frac{\partial^2}{\partial k_1^2}+\frac{\partial^2}{\partial k_2^2}+...+\frac{\partial^2}{\partial k_n^2}-1\right)k^{2b}{}_1F_2(a;b+\tfrac{1}{2}n,b+1;\tfrac{1}{4}k^2)=\frac{(b+\frac{1}{2}n-1)!b!}{(a-1)!}$$

$$\times\left[\frac{(a-1)!4k^{2b-2}}{(b+\frac{1}{2}n-2)!(b-1)!}+\sum_{s=0}^{\infty}\frac{(a+s-2)!(a-1)k^{2b+2s-2}}{4^{s-1}(b+\frac{1}{2}n-2+s)!(b+s-1)!s!}\right]$$

from which is evident that

$$\left(\frac{\partial^2}{\partial k_1^2}+\frac{\partial^2}{\partial k_2^2}+...+\frac{\partial^2}{\partial k_n^2}-1\right)k^{2b}{}_1F_2(1;b+\tfrac{1}{2}n,b+1;\tfrac{1}{4}k^2)=(b+\tfrac{1}{2}n-1)4bk^{2b-2} \qquad (A4.4)$$

and

$$\left(\frac{\partial^2}{\partial k_1^2}+\frac{\partial^2}{\partial k_2^2}+...+\frac{\partial^2}{\partial k_n^2}-1\right)k^{2b}{}_1F_2(a;b+\tfrac{1}{2}n,b+1;\tfrac{1}{4}k^2)$$

$$=4b(b+\tfrac{1}{2}n-1)k^{2b-2}{}_1F_2(a-1;b+\tfrac{1}{2}n-1,b;\tfrac{1}{4}k^2),\quad a\neq 1.$$

Consequently, if $a\neq 1$ or 2, then since

$$\left(\frac{\partial^2}{\partial k_1^2}+\frac{\partial^2}{\partial k_2^2}+...+\frac{\partial^2}{\partial k_n^2}\right)k^q=q(q+n-2)k^{q-2}$$

for all q except those for which $q+n=2,0,-2,-4,...$,

$$\left(\frac{\partial^2}{\partial k_1^2}+\frac{\partial^2}{\partial k_2^2}+...+\frac{\partial^2}{\partial k_n^2}-1\right)k^{2b}{}_1F_2(a;b+\tfrac{1}{2}n,b+1;\tfrac{1}{4}k^2)$$

$$=4^2b(b-1)(b+\tfrac{1}{2}n-1)(b+\tfrac{1}{2}n-2)k^{2b-4}{}_1F_2(a-2;b+\tfrac{1}{2}n-2,b-1;\tfrac{1}{4}k^2)$$

where, in deriving this result, since it cannot be assumed that $b-1>0$, with $b=N-\frac{1}{2}q-\frac{1}{2}n$ we impose the condition $q=2m\ \ (m=0,1,2,...)$. Thus, using equation (A4.4) we can write

$$\left(\frac{\partial^2}{\partial k_1^2}+\frac{\partial^2}{\partial k_2^2}+...+\frac{\partial^2}{\partial k_n^2}-1\right)^N k^{2N-q-n}{}_1F_2(N;N-\tfrac{1}{2}q,N+1-\tfrac{1}{2}q-\tfrac{1}{2}n;\tfrac{1}{4}k^2)$$

$$=\left(\frac{\partial^2}{\partial k_1^2}+\frac{\partial^2}{\partial k_2^2}+...+\frac{\partial^2}{\partial k_n^2}-1\right)4^{N-1}\frac{(N-\frac{1}{2}q-\frac{1}{2}n)!(N-\frac{1}{2}q-1)!k^{-q-n+2}}{(-\frac{1}{2}q-\frac{1}{2}n+1)!(-\frac{1}{2}q)!}$$

$${}_1F_2(1;-\tfrac{1}{2}q+1,-\tfrac{1}{2}q-\tfrac{1}{2}n+2;\tfrac{1}{4}k^2)=\frac{(N-\frac{1}{2}q-\frac{1}{2}n)!(N-\frac{1}{2}q-1)!4^Nk^{-q-n}}{(-\frac{1}{2}q-\frac{1}{2}n)!(-\frac{1}{2}q-1)!}. \tag{A4.5}$$

Using equations (A4.5) and (A4.3) in equations (A4.1) and (A4.2) we find that

$$F(\mathbf{k})=\frac{(N-\frac{1}{2}q-\frac{1}{2}n)!(\frac{1}{2}q+\frac{1}{2}n-N-1)!}{(-\frac{1}{2}q-\frac{1}{2}n)!(-\frac{1}{2}q-1)!}2^{q+n}(-1)^N\pi^{n/2}k^{-q-n}.$$

Finally, using the formula

$$z!(-z)!=\frac{\pi z}{\sin\pi z}$$

we have

$$(N-\tfrac{1}{2}q-\tfrac{1}{2}n)!(\tfrac{1}{2}q+\tfrac{1}{2}n-N-1)!=\frac{\pi}{\sin\pi(\frac{1}{2}q+\frac{1}{2}n-N)}$$

$$=\frac{(-1)^N\pi}{\sin\frac{1}{2}\pi(q+n)}=(\tfrac{1}{2}q+\tfrac{1}{2}n-1)!(-\tfrac{1}{2}q-\tfrac{1}{2}n)!(-1)^N$$

so that

$$F(\mathbf{k})=\frac{(\frac{1}{2}q+\frac{1}{2}n-1)!}{(-\frac{1}{2}q-1)!}2^{q+n}\pi^{n/2}k^{-q-n}.$$

We can write this result using the Gamma function notation where

$$m!=\Gamma(m+1)=\int_0^\infty t^m\exp(-pt)dt$$

which generalizes to values of m which are non-integer. Then,

$$F(\mathbf{k})=\frac{\Gamma\left(\frac{q+n}{2}\right)}{\Gamma\left(-\frac{q}{2}\right)}2^{q+n}\pi^{n/2}k^{-q-n}.$$

Hence, in the case when $n = 1$,

$$F(k) = \mathcal{F}_1[r^q] = \frac{\Gamma\left(\frac{1+q}{2}\right)}{\Gamma\left(-\frac{q}{2}\right)} 2^{1+q}\sqrt{\pi}k^{-q-1}$$

or

$$\mathcal{F}_1\left[\frac{1}{r^{1-q}}\right] = 2^q\sqrt{\pi}\frac{\Gamma\left(\frac{q}{2}\right)}{\Gamma\left(\frac{1-q}{2}\right)}\frac{1}{k^q}$$

and thus,

$$\mathcal{F}_1^{-1}\left[\frac{1}{(ik)^q}\right] = \frac{\alpha_1(q)}{r^{1-q}}$$

where

$$\alpha_1(q) = \frac{1}{(2i)^q\sqrt{\pi}}\frac{\Gamma\left(\frac{1-q}{2}\right)}{\Gamma\left(\frac{q}{2}\right)}.$$

For $n = 2$

$$\mathcal{F}_2\left[r^q\right] = \frac{\Gamma\left(\frac{q+2}{2}\right)}{\Gamma\left(-\frac{q}{2}\right)} 2^{q+2}\pi k^{-q-2}$$

or

$$\mathcal{F}_2\left[\frac{1}{r^{2-q}}\right] = 2^q\pi\frac{\Gamma\left(\frac{q}{2}\right)}{\Gamma\left(1-\frac{q}{2}\right)}\frac{1}{k^q}$$

and hence,

$$\mathcal{F}_2^{-1}\left[\frac{1}{(ik)^q}\right] = \frac{\alpha_2(q)}{r^{2-q}}$$

where

$$\alpha_2(q) = \frac{1}{(2i)^q\pi}\frac{\Gamma\left(1-\frac{q}{2}\right)}{\Gamma\left(\frac{q}{2}\right)}.$$

Thus, in general, ignoring scaling by $\alpha_1(q), \alpha_2(q), \alpha_3(q), ...$,

$$\mathcal{F}_n^{-1}\left[\frac{1}{(ik)^q}\right] \sim \frac{1}{r^{n-q}}, \quad n = 1, 2, 3, ...$$

YHTEENVETO (SUMMARY IN FINNISH)

Sähkömagneettinen sirontateoria on oleellinen, jotta voidaan ymmärtää vuorovaikutusta sähkömagneettisten aaltojen ja epähomogenisten dielektristen materiaalien välillä. Teoria avaa teknisen tiedon koskien suurta määrää sähkömagneettisia järjestelmiä, esimerkiksi optiikasta radio- ja mikroaaltokuvannukseen. Tarkkojen sirontamallien kehittäminen on erityisen tärkeätä kuvanymmärtämisen alalla ja tulkittaessa sirontatapausten synnyttämiä sähkömagneettisia signaaleja. Tätä päämäärää varten on olemassa joukko menettelytapoja, joita voidaan käyttää. Suhteellisen yksinkertaisia geometrisia konfiguraatioita varten käytetään likiarvometodeja kehittämään muuntamista kohdetasosta (missä sirontatapaukset tapahtuvat) kuvatasolle (missä tapahtuu jonkinasteinen sirontakenttä). Yleisin likiarvo on heikko sirontalikiarvo, joka ei ota huomioon monien sirontavuorovaikutusten vaikutusta. Tämän väitöskirjan, jonka nimi on Sähkömagneettisen sironnan ja käänteisen sironnan ratkaisujen käyttäminen digitaalisten signaalien ja kuvien analyysissä ja prosessoinnissa, ensimmäinen osa tutkii tämän likiarvon käyttöä sähkömagneettisten kuvannusjärjestelmien mallinnuksessa. Seuraavaksi väitöstyössä tarkastellaan lähestymistapaa, joka perustuu voimakkaaseen sirontajärjestelmään, johon kuuluu sirontakentän autokorrelaatio kehitettävissä käänteisissä sirontaratkaisuissa. Kun sirontavuorovaikutukset tulevat enenevästi monimutkaisemmiksi (esim. monet sironnat satunnaisvälineissä), deterministisen sirontateorian sovellukset tulevat vaikeiksi käyttää käytännössä. Näinollen käänteinen sirontaongelma ei välttämättä tule hyvin esitetyksi. Tästä syystä tarkastellaan useita muita lähestymistapoja, jotka sisältävät tilastollisten mallien kehittämisen itse sirontakentälle mieluummin kuin sirontajalle. Väitöskirjassa tutkitaan diffuusion käyttöä, joka perustuu malleille ratkaista käänteinen sirontaongelma, kun esiintyy voimakkaita sirontaprosesseja, esim. monisirontaa satunnaisvälineistä. Seuraavaksi lähestymistapaa laajennetaan ja käsitellään välitapausta mallintamalla sirontaprosesseja käyttäen murtolukuista diffuusioyhtälöä. Lopuksi esitetään matalataajuinen sirontatoria, joka johtaa esitykseen, että valo ja muut korkeataajuinen sähkömagneettisten aaltojen kentät voidaan heikosti difrahoida (taivuttaa) matalataajuisella sirontakentällä. Tämä johtaa uuteen tulkintaan gravitaatiolinssistä, jota tutkitaan kysymyksen kautta, miksi näkyvässä spektrissä havaitut Einsteinin renkaat ovat sinisiä.